

PhD thesis

Plausible auditory augmentation of physical interaction

submitted by

Marian Weger

at the

Institute of Electronic Music and Acoustics,
University of Music and Performing Arts Graz, Austria.

Committee:

Prof. Robert Höldrich	University of Music and Performing Arts Graz, Austria
Dr. Thomas Hermann	Bielefeld University, Germany
Prof. Gerhard Eckel	University of Music and Performing Arts Graz, Austria



Graz, April 2022

Copyright © 2022 Marian Weger
mail@marianweger.com



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license.



Transparency notice:

Large parts of this document are also published in other form. Such references are indicated at the beginning of the respective chapters or sections. All publications that relate to this dissertation at the time of writing are listed at the end (p. 223).

Abstract

We live in a physical environment in which interactions with physical objects evoke sound. This auditory feedback conveys information on the involved objects and on the specific type of interaction. We constantly adapt to the auditory feedback, even unconsciously, while pursuing our everyday activities. The digital environment which is becoming increasingly important lacks this immediate connection. It thus becomes necessary to project the digital information into the physical world in a plausible and usable way. With the visual domain being already overloaded, we propose the use of auditory augmentation to provide a calm communication channel by adding augmented auditory feedback to physical objects or interactions. Auditory augmentation is investigated within this thesis in 6 different ways. (1) We present experimental platforms for invisible auditory augmentation of everyday objects as well as for exploring the limits of plausibility. (2) As even naive listeners are already skilled in the interpretation of sounds of physical origin, we propose a physical sound model to synthesize augmented auditory feedback that integrates seamlessly in the everyday acoustic environment. (3) We review how physical information is encoded in auditory feedback, and investigate what portion of it is actually perceived and interpreted by human listeners, with a focus on rectangular plates. (4) We present an algorithm that successfully identifies material, size, and shape from sound. (5) We introduce an algorithm for auditory contrast enhancement which makes certain sound characteristics more salient. (6) We explore what kinds of data, physical objects, and interactions are suitable for auditory augmentation, and how much information it allows to monitor in the periphery of attention. Conclusions are drawn based on several case studies of auditory augmentations. This thesis provides the theoretical foundations as well as practical solutions and guidelines for designers of future auditory augmentations.

Kurzfassung

In unserer physikalischen Welt werden durch Interaktionen mit physikalischen Objekten hörbare Klänge hervorgerufen. Dieses auditorische Feedback vermittelt Informationen über die beteiligten Objekte, sowie über die Art der Interaktion. Im Alltag nutzen wir diese Informationen sowohl bewusst, als auch unbewusst, um unsere Handlungen an die Umgebung anzupassen. Der an Bedeutung gewinnenden digitalen Umgebung fehlt dieser unmittelbare Zusammenhang. Digitale Informationen müssen daher unaufdringlich, auf plausible und gebrauchstaugliche Art und Weise in die physische Welt projiziert werden. Ergänzend zum ohnehin überladenen visuellen Bereich erweist sich die auditorische Augmentierung als vielversprechende Lösung. Dabei wird physikalischen Objekten oder Interaktionen erweitertes auditorisches Feedback aufgeprägt, welches als zusätzlicher Informationskanal dient. In dieser Arbeit wird auditorische Augmentierung auf 6 unterschiedliche Arten erforscht. (1) Es werden Experimentalplattformen vorgestellt, die eine unsichtbare auditorische Augmentierung von Alltagsgegenständen ermöglichen, um in Folge die Grenzen der Plausibilität auszuloten. (2) Wir präsentieren ein physikalisches Modell zur Synthese von Klängen, die sich nahtlos in die gewohnte akustische Umgebung einpassen und ohne Training, auf Basis von Erfahrungen aus dem Alltag, sinnvoll interpretiert werden können. (3) Am Beispiel von rechteckigen Platten gehen wir der Frage nach, wie physikalische Informationen in deren Klang kodiert sind, und untersuchen, welcher Anteil davon tatsächlich vom Hörer ausgewertet wird. (4) Wir stellen einen Algorithmus vor, der Material und Abmessungen von rechteckigen Platten anhand deren Klang ermittelt. (5) Wir etablieren Methoden zur akustischen Kontrastverstärkung, mit dem Ziel, relevante Klangeigenschaften besser wahrnehmbar zu machen. (6) Anhand von Fallstudien wird untersucht, welche Arten von Daten, physikalischen Objekten und Interaktionen sich für die auditorische Augmentierung eignen, und wie viel Information damit am Rande der Aufmerksamkeitsschwelle vermittelt werden kann. Diese Arbeit liefert sowohl die theoretischen Grundlagen als auch praktische Lösungen und Empfehlungen für die Entwicklung zukünftiger auditorischer Augmentierungen.

Erklärung

Name: Marian Weger

Matrikelnummer: 00673093

Studienrichtung: 094 PhD-Studium (Doctor of Philosophy), 750 Elektrotechnik-Toningenieur

Hiermit bestätige ich, dass mir der *Leitfaden für schriftliche Arbeiten an der KUG* bekannt ist und ich diese Richtlinien eingehalten habe.

Graz, den _____
(Datum)

(Unterschrift des Verfassers)

Acknowledgments

This work would not have been possible without the support of some great people and institutions. I would like to express my appreciation.

To Robert Höldrich.

To Thomas Hermann.

To Gerhard Eckel.

To my family.

To my colleagues at the IEM, especially Katharina Groß-Vogt.

To my (former) students, especially Martin Czuka, Michael Aurenhammer, and Iason Svoronos-Kanavas.

To the members of the Ambient Intelligence Group at CITEC, who warmly welcomed me during my research stays in Bielefeld.

To _____.¹

This research was partially supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

¹If you think you should also appear in this list, please enter your name.

Contents

Abstract	V
Kurzfassung	VII
Acknowledgments	XI
Contents	XIII
1 Introduction	1
1.1 Plausible auditory feedback	3
1.2 Usable auditory feedback	4
1.3 Interaction with the digital environment	5
1.4 Activities, actions, and behavior	6
1.5 Natural users and reality-based interaction	8
1.6 Auditory augmentation and auditory display	9
1.7 What sounds plausible?	14
1.8 Conclusions	18
Bibliography	18
2 The psychophysics of auditory feedback: an annotated bibliography	23
2.1 Source identification, causal uncertainty, and informational masking	23
2.2 Auditory perception of physical parameters	26
2.2.1 Material category	26
2.2.2 Action	30
2.2.3 Material properties	30
2.2.4 Force of impact	32
2.2.5 Surface roughness	32
2.2.6 Pressure and speed of rubbing	32
2.2.7 Hardness	33
2.2.8 Impact position	33
2.2.9 Point-shaped external damping	34
2.2.10 Size	34
2.2.11 Shape of rigid objects	35
2.2.12 Shape and volume of cavities	36
2.2.13 Distance	37
2.2.14 Hallowness	37
2.2.15 Boundary condition	37
2.2.16 Veracity: real vs. synthetic	38
2.2.17 Summary	38
2.3 Perceptual resolution of sound parameters	39
2.3.1 Decay times	39
2.3.2 Amplitudes	39
2.3.3 Missing partials	39

2.3.4	Base frequency	40
2.3.5	Frequency ratios / intervals	40
2.3.6	Modal density	40
2.3.7	Upper cutoff frequency / low-pass filtering	40
2.3.8	Lower cutoff frequency / high-pass filtering	41
2.3.9	Summary	41
2.4	Robotic auditory perception of physical parameters	42
2.4.1	Material	42
2.4.2	Shape	44
2.4.3	Action	44
2.4.4	Summary	44
2.5	Multisensory feedback and the integration of auditory, visual, and haptic information	45
2.5.1	Latency and temporal resolution	45
2.5.2	Auditory-haptic feedback	48
2.5.2.1	Influence on walking style	48
2.5.2.2	Material category	49
2.5.2.3	Material properties	50
2.5.2.4	Surface texture	51
2.5.2.5	Subjective quality	52
2.5.2.6	Summary	52
2.5.3	Auditory-visual feedback	52
2.5.3.1	Material category	52
2.5.3.2	Veracity and plausibility	53
2.5.3.3	Influence on grasping actions	54
2.5.3.4	Summary	55
2.5.4	Auditory-visual-haptic feedback	55
2.5.4.1	Material properties	55
2.5.4.2	Naturalness, usability, and pleasantness	57
2.5.4.3	Influence on typing performance	57
2.5.4.4	Summary	58
3	Physical modeling for plausible auditory augmentation	59
3.1	Acoustics of simple systems	59
3.1.1	Undamped harmonic oscillator	59
3.1.2	Damped harmonic oscillator	60
3.1.3	Driven harmonic oscillator	61
3.1.4	Damping parameters and their relationships	61
3.1.5	Coupled oscillators and modes	62
3.2	Acoustics of continuous systems: bars and plates	63
3.2.1	Free vibrations of thin bars	63
3.2.2	Free vibrations of thin plates	64
3.2.3	The impulse response of a rectangular plate	65
3.2.4	Mode shapes	66
3.2.5	Undamped natural frequencies	66
3.2.6	Damping	67
3.2.6.1	Radiation damping	68
3.2.6.2	Thermoelastic damping	69
3.2.6.3	Viscoelastic damping	70
3.2.6.4	Viscous damping	72
3.2.6.5	Complete damping model	72
3.2.7	Modal weights due to the position of contact	73
3.2.8	Hertz' law of contact	73
3.2.9	Indentation hardness	74

3.2.10	Radiation efficiency	76
3.3	Modal analysis and synthesis	77
3.3.1	Modal analysis	77
3.3.2	Modal synthesis	77
3.3.2.1	Finite difference approximation	78
3.3.2.2	Simple resonator	78
3.3.2.3	Smith-Angell resonator	79
3.3.2.4	Validation	79
	Bibliography	79
4	“AltAR/table”: an experimental platform for plausible auditory augmentation	83
4.1	Hardware platform	83
4.2	Equalization	85
4.3	Measurements	88
4.4	Spatialization and low-frequency extension	90
4.5	Subtractive modal synthesis model	91
4.6	Noise and feedback control	91
4.6.1	Anti-resonator	92
4.6.2	Frequency shifting	94
4.6.3	Feedback cancellation	95
4.6.4	Notch filters	96
4.7	Tracking of contact position and hand damping	96
4.7.1	Markerless position tracking	96
4.7.2	Marker-based position tracking	97
4.7.3	Pressure sensing	98
4.8	Discussion	98
	Bibliography	98
5	Auditory perception and information capacity of rectangular plates	101
5.1	An algorithm for robotic perception of material and dimensions of rectangular plates	101
5.1.1	Measuring sound parameters: amplitudes, frequencies, decay factors	102
5.1.1.1	Natural frequencies	102
5.1.1.2	Modal weights and decay factors	103
5.1.1.3	Critical frequency of radiation damping	103
5.1.2	Estimating physical parameters from sound parameters	104
5.1.2.1	Internal damping and material	104
5.1.2.2	Dimensions and boundary conditions	104
5.1.3	Case studies	107
5.1.4	Discussion	108
5.2	Multisensory discrimination of size and aspect ratio	109
5.2.1	Stimuli	110
5.2.2	Apparatus	110
5.2.3	Procedure	111
5.2.4	Participants	112
5.2.5	Results	112
5.2.5.1	Outliers and individual participants’ accuracy	112
5.2.5.2	Confusion between parameter dimensions	113
5.2.5.3	Discrimination between main levels of length and aspect ratio	113
5.2.5.4	Direction and amount of a parameter change	114
5.2.5.5	Estimated vs. true values	114
5.2.5.6	Number of discriminable levels	114
5.2.5.7	Confusion between individual shapes	116
5.2.5.8	Individual participants’ interaction strategies	117

5.2.6	Discussion: why is it so difficult?	119
5.3	Auditory discrimination of material and aspect ratio	120
5.3.1	Stimuli	121
5.3.2	Apparatus and procedure	123
5.3.3	Participants	123
5.3.4	Results	124
5.3.4.1	Overview	124
5.3.4.2	Confusion between non-metals and metals	125
5.3.4.3	Confusion between rigidities	126
5.3.4.4	Confusion between aspect ratios	127
5.3.4.5	Number of discriminable levels	127
5.3.5	Discussion and conclusions	128
5.4	The information capacity of multidimensional auditory displays	129
	Bibliography	131
6	“Schrödinger’s box”: an experimental platform for implausible auditory augmentation	133
6.1	The plausibility of auditory feedback	134
6.2	A taxonomy of sounds for a black box	135
6.3	Hardware platform	136
6.4	Real-time onset detection revisited	137
6.4.1	Onset time: predict the future and undo the past	138
6.4.2	Impact force and spatial location: extract features before they emerge	140
6.5	Sample playback	140
6.6	Conclusions and outlook: exploring the limits of plausibility	141
	Bibliography	141
7	Prospects and limits of auditory augmentations	145
7.1	An interdisciplinary workshop on auditory augmentation	145
7.2	Datasets and sonification platforms	146
7.2.0.1	Datasets	146
7.2.0.2	Sonification platforms	146
7.3	Prototypes	147
7.3.1	“Writing resonances”	147
7.3.2	“Exploration table”	148
7.3.3	“Sonic floor plan”	149
7.3.4	“Smart kettle”	149
7.3.5	“Standby door”	150
7.3.6	“3D gestural mouse”	150
7.3.7	“Hob assistant”	151
7.3.8	“Kitchen sounds”	151
7.3.9	“Interleave”	151
7.4	Discussion	152
7.4.1	On the peculiarity of sound in augmented reality	152
7.4.2	So what is auditory augmentation?	152
7.4.3	The relationship between sound, data, and augmented object	153
7.4.4	Perceptual implications of auditory augmentation	154
7.4.5	Why auditory augmentation?	156
	Bibliography	156
8	Case studies of auditory augmentations	159
8.1	“Mondrian table”: display of spatial information by auditory augmentation	159
8.1.1	Augmented graphic tablet	160
8.1.2	Auditory coloring book	163

8.1.3	Discussion	164
8.2	“PilotKitchen”: display of electric load by virtual room acoustics	164
8.2.1	Apparatus	165
8.2.2	Electric load in the kitchen	166
8.2.3	Data processing	166
8.2.4	Sonification design	168
8.2.5	Evaluation	168
	8.2.5.1 Perceptibility	168
	8.2.5.2 Contextualization	169
8.2.6	Conclusions	170
8.3	“RadioReverb”: room reverberation as ambient communication channel	171
8.3.1	Participants and their main task	171
8.3.2	Stimuli: artificial room reverberation	171
	8.3.2.1 A plausible range of reverberation	171
	8.3.2.2 Binaural rendering of virtual room acoustics	172
	8.3.2.3 Discriminable levels of reverberation	172
8.3.3	Apparatus: a mobile experiment app	173
8.3.4	Procedure	173
	8.3.4.1 Training and main experiment	173
	8.3.4.2 Post-hoc survey	174
8.3.5	Results	174
8.3.6	Discussion	177
8.4	“CardioScope”: an augmented stethoscope using ECG data	178
8.4.1	Electrocardiography and the stethoscope	178
8.4.2	Synchronous acquisition of ECG and PCG	178
8.4.3	ECG-informed auditory augmentation of heart sounds	180
8.4.4	Discussion	181
	Bibliography	181
9	Auditory contrast enhancement (ACE)	185
9.1	Real-time auditory contrast enhancement	186
9.1.1	Spectral contrast enhancement	187
	9.1.1.1 Spectral sharpening by lateral inhibition	190
	9.1.1.2 Spectral dynamics expansion by exponentiation	191
	9.1.1.3 Decay prolongation by envelope processing	191
	9.1.1.4 Vibrato expansion by frequency shifting	193
9.1.2	Temporal contrast enhancement	194
9.1.3	Discussion	195
9.1.4	Conclusions	196
9.2	Plausibility of enhanced auditory feedback	196
9.2.1	Stimuli	196
9.2.2	Participants, apparatus, and procedure	197
9.2.3	Results	199
9.2.4	Discussion	199
9.3	The ACE hear-through system	200
9.4	Conclusions and outlook	201
	Bibliography	201
10	General conclusions and outlook	205
	Bibliography	208
A	Appendix to the physical model	209
A.1	Approximate characteristic beam functions	209

Contents

A.2 Approximate frequency factors of rectangular thin plates	211
Bibliography	212
List of Figures	213
List of Tables	219
List of Supplementary Material	221
Related publications	223

1. Introduction

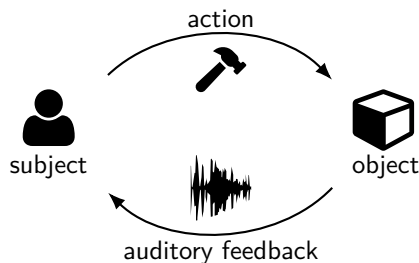


Figure 1.1.: The interaction loop with auditory feedback.

As we live in a physical world, we constantly interact with other physical objects: living beings, the floor we are walking on, objects we take in our hands, or a table where such objects may be positioned. Even non-living objects may interact with each other, as we experience from the forces of nature. Any of such physical interactions typically evoke sounds, as long as some medium such as air or water transports the physical vibration to our ears. To a large extent, this sound is highly predictable. It depends, for example, on the physical properties of the involved objects, such as material, shape, or size, and also on the type of interaction, such as bouncing, hitting, or scratching. A major part of our everyday acoustic environment results from our own actions. In this case, we refer to it as auditory feedback (see Fig. 1.1).

Even if we are not always fully aware of it, we use this auditory feedback all the time while pursuing our daily activities, because it conveys information about our physical environment. For example, we shake objects or knock on them to retrieve information concerning their contents. We shake birthday presents to guess what we are receiving. We shake food packaging to estimate how much milk/coffee/etc. is left. We knock on barrels to estimate their filling levels. We even knock on doors of houses when we want to know if someone is there. All these were only conscious actions, representing active listening practices.

Most of the time, however, we use the information that is obtained from auditory feedback without even noticing it: to adjust subconscious processes

together with information from other senses such as vision or touch. We adapt our walking style to the sound of the floor—for example, when adjusting to the height of snow or undergrowth in unknown terrain. We integrate auditory information when pouring water (it is easier with sound!). In the dark, we navigate through the house by using our ears (try it with blocked ears!). When using mechanic tools or machines, their auditory feedback is often the only way to obtain information on their current status. For example, we change gears of the car based on the sound of the motor, or we adjust our fingers to the computer keyboard based on the mechanical (or nowadays simulated) sound. For such tools, the auditory feedback is critical to close the interaction loop and thus allow the tool to become an extension to our body.

We surely interact also with non-physical entities such as ideas or information from our digital environment. By themselves, however, these provide no auditory feedback. That seems to be a general problem of the digital environment: it needs to be projected in our physical environment, in order to materialize in the form of a stimulus that can be perceived and interpreted by our senses. The first choice is usually to present the digital information through a visual display, e.g., of a smartphone. For many data, especially the highly complex information that is shared in social media, this works pretty good. If the information, however, is tightly coupled to time, then it is sometimes better conveyed through sound—by means of sonification. That is why alarms are most often displayed by sound: anyone is able to perceive them at the right time, no matter what the person is currently doing. The temporal resolution of the ear generally outperforms that of vision, while vision excels in spatial resolution. If in addition to time, the information is tightly coupled to a physical object, then only sound can successfully close the interaction loop. For instance, most cars with combustion engine have a revolution counter in form of a visual display. It would require almost our full visual attention to read it while driving—the visual attention that we need for continuously monitoring the highly complex surroundings that only our eyes are able to capture.

1. Introduction

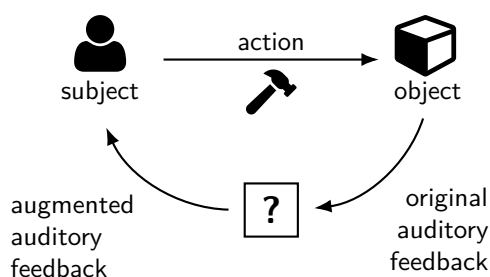


Figure 1.2.: The interaction loop with augmented auditory feedback.

The auditory feedback instead is able to convey the very simple information in the periphery of attention, with the high temporal resolution that is needed to turn the machine into an extension of our body.

The car example may be irrelevant in the future, but it leads us to a very popular example of sonification. The sound is only a by-product of an outdated technology that is already being replaced. Modern electric cars have no gears, and the motor emits almost no sound. The user receives no feedback at all on the machine's inner state, contrary to what was the case with combustion engines. All relevant information therefore needs to be projected, for example, by sound, so that the user or pedestrians in the vicinity are able to perceive it. Likewise, any other originally silent physical object may be augmented by artificial auditory feedback, in order to convey information by sound. For example, a water jar may carefully remind a user who forgot to drink (Groß-Vogt 2020). Even if it the physical object or action already provides useful original auditory feedback, additional sound may be added, so that even more information is conveyed. An electric drill may, for example, sonify the tilt angle by additional artificial auditory feedback, so that the user is able to drill in the desired angle without a drilling rig (Großhauser and Hermann 2010). If the original auditory feedback is modulated or augmented by additional sound, we speak of *augmented auditory feedback*: the artificially modified sonic reaction to physical interaction. This is visualized in Fig. 1.2.

More generally, we speak of *auditory augmentation* if a physical object or its sound is augmented by sound for the purpose of sonification, i.e., to convey additional information.

Augmented auditory feedback could be exploited in three different ways. First, the signal-to-noise ratio may be improved in order to help communication of (a) the involved objects' physical properties

such as material or spatial dimensions, and (b) the interaction type such as tapping or scratching. Both may facilitate specific activities. This is what we call auditory contrast enhancement, and what is addressed in Ch. 9.

Second, specific physical properties of a physical object could be modified perceptually, in order to induce a certain change in user behavior. Apart from the aforementioned footstep sounds, successful behavioral change through sound has been shown for hand tapping (Furfaro et al. 2013; Furfaro et al. 2015), as well as for grasping (Castiello et al. 2010; Sedda et al. 2011).

Finally, augmented auditory feedback creates a new communication channel that can also be used for sonification of data that are completely unrelated to both object and interaction, e.g., for continuous monitoring as a secondary task. The sonification is thus naturally and seamlessly fitted into the everyday acoustic environment.

A major goal of this thesis is to explore the exploitability of augmented auditory feedback as communication channel. In information theory, an important characteristic of a given channel is its channel capacity, i.e., the highest rate at which information can be reliably transmitted. In our case, it is limited by at least two factors: the *plausibility* and the *usability* of augmented auditory feedback. Both will be discussed in detail in Sec. 1.1 and 1.2. For a specific physical interaction, we make four assumptions:

- (1) There exists a manifold of sounds which serve as *plausible auditory feedback*. Its borders define the *plausibility range*.
- (2) There exists a manifold of sounds which serve as *usable auditory feedback*, i.e., sounds that help to perform specific actions. Its borders define the *usability range*.
- (3) The two manifolds of plausible and usable sounds overlap. We define this overlap region as the manifold of *alternative auditory feedbacks*.
- (4) It is possible to discriminate between different alternative auditory feedbacks.

If all these assumptions are met, we conclude that it is possible to convey additional information through (more or less subtle) sound changes within the intersection of plausible and usable auditory feedbacks. Contrary to the oversimplified depiction in Fig. 1.3, plausibility and usability are no binary

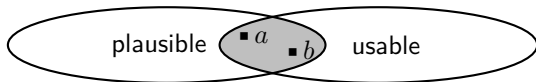


Figure 1.3.: Sets of plausible and usable variants of auditory feedback for a specific physical interaction. Points *a* and *b* represent two discriminable but still plausible and usable sounds.

states, but rather continuous qualities whose actual characteristics are unknown at this point. The plausibility range and usability range are not meant as strict borders but describe the sets of sounds that a majority of persons perceives as plausible or usable, respectively—for the given physical object(s) and interaction as well as for the specific situation and context.

Note that usability in this argumentation refers to the original physical interaction. Augmented auditory feedback which lacks relevant information from the original auditory feedback is considered to be less usable. Even if the net information capacity is not decreased, e.g., by adding different information, the original interaction is still assumed to be deteriorated.

On a pure physical level, all the known physical (fundamental) interactions or forces of nature are governed by the four basic forces (gravitational, electromagnetic, strong, and weak force). In the context of this thesis, we mainly refer to those physical interactions that can be described by classical mechanics: interactions between physical objects or between humans and physical objects. While physical objects usually include all objects with spatial location (Markosian 2000), we usually mean solid or rigid-body objects.




The remainder of this chapter is structured as follows. In Sec. 1.1 and 1.2, the concept of plausible and usable auditory feedback is investigated in more detail. Then, in Sec. 1.3, we will recapitulate the terminology of human–computer interaction (HCI), followed by the theoretical framework of activity theory in Sec. 1.4. Sections 1.5 and 1.6 review general concepts of HCI, auditory display, and interactive sonification. In Sec. 1.7 we will take a closer look on the meaning of plausibility and related theories.

The remaining chapters of this thesis constitute its major contributions. Chapter 2 provides a comprehensive literature review on auditory perception of physical sound sources. Chapter 3 contains an introduction to modal analysis and synthesis and describes a physical sound model of rectangular



Figure 1.4.: Two bells and their expected sound.

plates, targeting auditory augmentation. Chapter 4 documents an experimental hardware and software platform for the auditory augmentation of rigid surfaces. Chapter 5 introduces a novel algorithm for robotic perception as well as two listening experiments on human perception of material, size, and shape of rectangular plates. Chapter 6 presents a platform for exploring the (im-)plausibility of auditory feedback. Chapter 7 discusses the prospects and limits of auditory display, based on the results of an interdisciplinary prototyping workshop. Chapter 8 documents several case studies of auditory augmentations and their evaluations. Chapter 9 introduces the concept of auditory contrast enhancement (ACE) together with real-time capable algorithms. Chapter 10 finally recapitulates the major contributions of this thesis and provides conclusions as well as an outlook on future research.

This document links to several sources of supplementary materials:  audio examples,  video examples, and  source code. If you are using a digital copy, then just click on the individual icons whenever they appear alongside the text to access the additional content on the internet. A list of all supplementary material (including the corresponding links) is provided on p. 221.

1.1. Plausible auditory feedback

Sensory feedback is generally considered to be plausible if it is “conceptually consistent with what is known to have occurred in the past” (Connell and Keane 2006). In particular, “a highly plausible scenario is one that fits prior knowledge well: with many different sources of corroboration, without complexity of explanation, and with minimal conjecture”. In other words, something is plausible if our expectations are met and if the individual feedback from different sensory modalities is in agreement with each other. For example, we would generally assume that a small object has higher pitch than a large object, as depicted in Fig. 1.4.

With increasing perceptual dominance of auditory augmentation, the resulting auditory feedback influences user perception, emotion, and behavior (Furfaro et al. 2015). For example, auditory cues

1. Introduction

influence the haptic perception of virtual textures (Serafin et al. 2007). Likewise, perception of material properties (e.g., hard/soft, rough/smooth) is strongly influenced by auditory cues (Martín et al. 2015). As the perceptual plausibility depends on the congruency between different modalities such as haptic, visual, or auditory information, it has no meaning for the unisensory case of auditory feedback alone. Perceptual congruency and therefore plausibility is high if the information of the different modalities combined, i.e., the combination of different stimuli, matches the pattern we learned through natural interactions with our physical environment. It is therefore hypothesized that perceptual plausibility increases with increasing congruency (agreement) between cues (information) from different sensory modalities (information channels).

It must be considered that people are already accustomed to manipulated visual feedback, but generally have less experience (or none at all) in augmented auditory or tactile feedback. It is common knowledge that a physical object's inner structure can be concealed, e.g., through painted surfaces. This supports the assumption that there exist at least several interchangeably plausible visual representations of a physical object.

We argue that if an auditory augmentation alters the perception of only such physical properties that are hidden behind the surface finishing (e.g., lacquer or laminate), i.e., physical properties that are not conveyed through vision or haptics, then the auditory augmentation cannot lead to incongruent sensory information. The augmented physical properties concerning the inner structure of physical objects may include, for example, material category, density, hollowness, and spatial volume, as well as boundary condition (e.g., free or clamped) and coupling to other physical objects. Their individual exploitability for the purpose of sonification depends on their perceptual resolution. The literature review on auditory perception in Ch. 2 will shed light on that.

If the above material properties have a perceivable effect on auditory feedback, their modulation through auditory augmentation is supposed to provide a reliable communication channel. The resulting augmented auditory feedback is assumed to stay within the plausibility range if augmented in a physically meaningful and feasible way, i.e., if the illusory physical properties can be explained without effort. The meaning of plausibility will be further discussed in Sec. 1.7. For now we rather proceed to the second limiting factor of the information capacity of

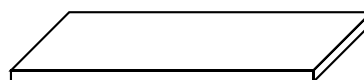


Figure 1.5.: An even, horizontal, rigid, and stationary surface.

auditory augmentations: usability.

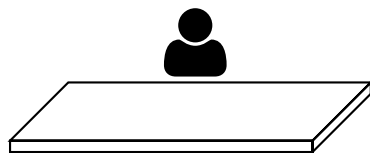
1.2. Usable auditory feedback

usability “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” — ISO 9241-210 (ISO 2009)

Auditory feedback in general has a positive effect on task performance and motor learning (Sigrist et al. 2013). Sigrist et al. provided design criteria to successful visual, auditory, haptic, and multisensory feedback in the context of motor learning. They argue that a positive effect of auditory feedback observed in isolation may completely vanish in the presence of feedback in other sensory modalities. For example, auditory feedback improves the typing performance on a computer keyboard if no additional haptic feedback was present; however, the influence of haptics is much stronger (Ma, Zhaoyuan et al. 2015).

In any case, what makes auditory feedback usable is the information it carries about the performed action and about the physical objects it affects. All information that is encoded in the auditory feedback but not needed for the given task with the physical object in the given context, is therefore assumed to be irrelevant for usability. We define *relevant physical properties* as the properties of a physical object which influence its usability. Likewise, *irrelevant physical properties* are the properties of a physical object which do not affect its usability.

Irrelevant physical properties are therefore considered as possible candidates for usability-independent auditory augmentation. Note that auditory augmentation of irrelevant physical properties may mask the perception of relevant physical properties — either auditorily or through informational masking effects, e.g., by adding additional disturbance or stress. Furthermore, the relevance of specific physical properties of the same physical object diverge for different actions and context.



“What about the legs?”

Figure 1.6.: A person sitting at a table.

As an example, we examine a relatively simple category of physical objects: an even, horizontal, rigid, and stationary surface (see Fig. 1.5). Such a surface usually appears on tables (see Fig. 1.6), cupboards, bookshelves, etc. Due to its affordances (Norman 2013, pp. 10–13), it is primarily used for putting things on top, moving these things around, and manipulating them (writing, cooking, etc.), but also manual interaction is possible (hitting, scratching, tapping, etc.).

For the observed surface, we consider three relevant physical properties. Hardness influences how we put fragile things on top or how we interact with our hands. Roughness influences how objects or fingers can be moved (see Fig. 1.7). Sturdiness is relevant, as a fragile table might break while positioning heavy things. We thereby do not want to modulate these but rather leave them unchanged in order to preserve usability.

Similarly, physical properties that we consider as irrelevant include spatial volume, hollowness, underlying material category, and boundary conditions— if not in conflict with relevant physical properties. These relate mainly to the non-visible part of the surface, under the visible texture layer. In consequence, their perception may be securely altered through auditory augmentation within the plausibility range.

Note that not all physical objects incorporate a specified purpose or intended use. Nevertheless, every physical object has affordances. The goal of usable auditory augmentation is to preserve these in the best possible way while adding new affordances such as exploratory data analysis through manual interaction.

1.3. Interaction with the digital environment

Physical objects also include machinery or computers. In this case, we usually do not directly interact with the actual physical process, but rather with an interface that somehow translates between

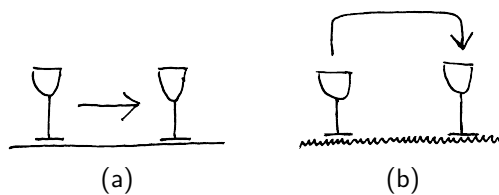


Figure 1.7.: The roughness of a table may influence, how objects are moved. Sliding on a smooth surface (a), lifting on a rough surface (b).

human and machine. As the vocabulary of human–computer interaction (HCI) is sometimes confusing or even misinterpreted, we will now briefly review it.

It all starts with a human *user action*. In a machine-centered view, i.e., from the perspective of the machine, this corresponds to an input. Such a user action can be anything, from pressing a button to yawning. In this context, we distinguish between active and passive *action modalities*. “For inputs, active modalities are used by the user to issue a command to the computer (e.g., a voice command or a gesture recognized by a camera). Passive modalities refer to information that is not explicitly expressed by the user, but automatically captured for enhancing the execution of a task” (Nigay and Coutaz 1993).

The thereby transmitted information (in this case from human to computer) creates an *information channel*. “In typical HCI usage, a channel describes an interaction technique that utilizes a particular combination of user ability and device capability (such as the keyboard for inputting text, a mouse for pointing or selecting, or a 3D sensor used for gesture recognition). In this view, the following are all channels: text (which may use multiple modalities when typing in text or reading text on a monitor), sound, speech recognition, images/video, and mouse pointing and clicking” (Turk 2014).

An interactive system may switch between different input or output channels depending on its *mode of interaction*. “Mode refers to a state that determines the way information is interpreted to extract or convey meaning” (Nigay and Coutaz 1993).

From a user-centered perspective, we humans receive stimuli from different *sensory modalities*, referring to our human senses such as audition, vision, touch, olfaction, or gustation. Touch or haptics, are umbrella terms that include cutaneous, kinesthetic, and tactile perception.

The commonly used terms *multimodality* and *multimodal* are obviously ambiguous, as they may

1. Introduction

refer to either modality (action modality, sensory modality) or mode. We therefore try to avoid them wherever possible or prefer *multisensory* when referring to a combination of multiple sensory modalities. For *cross-modal* effects which describe interactions between different sensory modalities, however, we need to make an exception.

A stimulus in one of the sensory modalities leads to a *sensation*: “the process by which stimulation of a sensory receptor gives rise to neural impulses that result in an experience, or awareness, of conditions inside or outside the body” (Gerrig 2013, p. 80).

Consequently, *perception* is “the process or result of becoming aware of objects, relationships, and events by means of the senses, which includes such activities as recognizing, observing, and discriminating. These activities enable organisms to organize and interpret the stimuli received into meaningful knowledge and to act in a coordinated manner” (VandenBos 2015, p. 775).

A user *interface* finally comprises “all components of an interactive system (software or hardware) that provide information and controls for the user to accomplish specific tasks with the interactive system” (ISO 2009).

Multimodal interfaces or multimodal systems “process two or more combined user input modes—such as speech, pen, touch, manual gestures, gaze, and head and body movements—in a coordinated manner with multimedia system output” (Jacko 2012). “Both multimedia and multimodal systems use multiple communication channels. But in addition, a multimodal system is able to automatically model the content of the information at a high level of abstraction. A multimodal system strives for meaning” (Nigay and Coutaz 1993).

1.4. Activities, actions, and behavior

Within the HCI community, human actions are often explained by adopting principles that build on Sergei Rubinstein’s psychological activity theory from the 1930s. According to Kaptelinin (1995), activity theory can be structured into some basic principles which are not to be interpreted as isolated ideas but are in fact closely interrelated.

Most fundamentally, activity theory discerns between three types of processes (activities, actions, and operations) based on a hierarchical model (see Fig. 1.8). These three levels are oriented to different

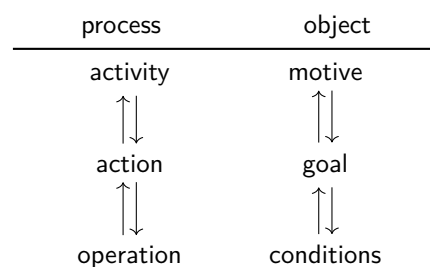


Figure 1.8.: Hierarchical model of processes (activities, actions, and operations) and their corresponding objects (motives, goals, conditions) in activity theory, after Kuutti (1995).

kinds of objects. Objects, in this scope, refer to the non-physical meaning, in the sense of aims or intentions. As depicted in Fig. 1.8, activities are oriented to motives, actions are oriented to goals, and operations adapt to conditions. But let us start with some practical examples.

Activities may comprise, for example, building a house or going on vacation. They are oriented to more general *motives* such as having a roof over one’s head or taking a rest after a busy period of work. Motives are impelling by themselves. They are objects or ideals that satisfy a certain need. When motives are frustrated, people usually get frustrated too, which might result in rather unpredictable behavior. (Kaptelinin 1995)

Activities usually involve a number of functionally subordinated processes: *actions*. Actions are always directed to specific (conscious) *goals*. If keeping with the above examples, a corresponding action would be to hammer a nail through a strip of wood, with the goal to attach it to a wooden beam. Similarly, the vacationer could drive the car from home to the hotel at a given route. Kaptelinin states that if a goal is frustrated, then people adapt to the new circumstances and set a new goal. This usually doesn’t imply much effort or negative emotions. In the examples, one might buy better nails, or change the planned route. Kuutti (1995) adds that actions typically follow an *orientation* phase in which the action is planned in the consciousness using a mental model. In general, the quality of this model is critical for the success of the performed action.

Actions comprise chains of (unconscious) operations. According to Kuutti, these are well-defined routines that are performed subconsciously in response to the *conditions* which are faced while an action is executed. Each operation starts as a con-

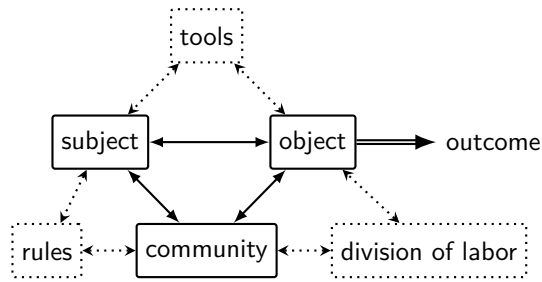


Figure 1.9.: Basic structure of an activity, after Kuutti (1995).

scious action, including its orientation phase and execution phase. However, the orientation phase fades out and the action collapses to an operation as soon as the corresponding mental model has reached a certain quality and the action has been practiced sufficiently long. Kaptelinin adds that people usually not even notice it when operations are frustrated, and unconsciously adapt to the new situation (e.g., if familiar conditions are changed). Examples for operations include the individual processes while hammering a nail (holding the nail, aiming, swinging the hammer, etc.), or the subconscious operations while driving a car: keeping a certain velocity and staying within the lane through (micro-)adjustments of pedals and steering wheel, changing gears, etc.

According to Kuutti, an activity can also be structured based on a systemic model, as depicted in Fig. 1.9. The individual object of an activity distinguishes it from other activities. The existence of an activity is only motivated by its *outcome*: the object transforms into an outcome as soon as it is accomplished or frustrated. The object is not necessarily a material thing but may also describe a more abstract and less tangible plan or idea, under the premise that it can be shared with the *community*. The community involves all participants of an activity. They may manipulate and transform the object and also contribute to it through their own actions.

The model in Fig. 1.9 visualizes the mutual relationships between the individual subject, the community (other subjects who are part of the activity), and the object. These relationships are mediated through different concepts.

Different types of *tools* may be used to accomplish a certain object (e.g., the hammer or the car in the above examples). Tools thus mediate the relationship between subject and object. They are not necessarily physical objects, but may also be

abstract tools for thinking. (Kuutti 1995)

Individual subjects interact with others based on certain *rules* that apply within this specific community (e.g., conventions or laws within a society). Rules therefore mediate the relationship between subject and community. They cover any type of explicit or implicit norms, conventions, or social relations that apply within the community. (Kuutti 1995)

The members of a certain activity usually perform different actions that are part of it, i.e., the activity is split into several parts which are assigned to the involved subjects either explicitly or implicitly. The principle of *division of labor* thus mediates the relationship between community and object. It describes the organization of the community concerning the shared activity. (Kuutti 1995)

Activity theory further includes the concept of *internalization*, which means that subject and object influence and even transform each other. While it is self-evident that the subject transforms the object, also “the properties of the object penetrate into the subject and transform him or her” (Kuutti 1995). The opposite process, *externalization*, applies if “mental processes manifest themselves in external actions performed by a person, so they can be verified and corrected, if necessary” (Kaptelinin 1995). The internal and external side of an activity can only exist together (Kuutti 1995).

Activities are further regarded as dynamic entities which always change and develop over time, at all levels within the hierarchic model from Fig. 1.8. The primary source of *development* is attributed to so-called *contradictions*. These occur in case of external influences which may cause imbalances within or between activities. Contradictions indicate an unfit within or between elements or between different development phases of an activity, or between activities. These usually manifest themselves in the form of problems, ruptures, breakdowns, etc. (Kuutti 1995)

Kaptelinin argues towards the application of activity theory as theoretical basis for human–computer interaction. In extension to cognitive psychology, it may account for social interactions, cultural factors, as well as aspects concerning development and higher-level goals and values, while at the same time incorporating experimental results and methods from cognitive psychology. (Kaptelinin 1995, p. 57)

In this sense, an *interface* is regarded as a link which allows to integrate a computer tool within the structure of human activities. Within activity

theory, the underlying mechanisms of an interface are assumed to form a functional organ, meaning that computer applications become the extensions of pre-computer human abilities. Concerning tool mediation, there are two interfaces to be considered: the human–computer interface and the computer–environment interface. (Kaptelinin 1995, p. 56)

1.5. Natural users and reality-based interaction

In the search for better interfaces, i.e., such that integrate better in our physical environment, different concepts and theories have been developed.

Tangible user interfaces (TUI) map digital data to physical real-world objects, so that a manipulation of the data is achieved by manual manipulation of the physical representations (Ullmer and Ishii 2000). Information is usually coupled to the physical artifacts by means of sound and video projection. A major achievement of tangible interfaces is that abstract data is actually made graspable. Even more importantly, the same data can be manipulated by multiple users at the same time, with the benefits and constraints of purely physical objects.

Dourish (2001) developed a new approach for interacting with computers: embodied interaction. According to Hartson and Pyla (2012, p. 328), “embodied interaction refers to the ability to involve one’s physical body in interaction with technology in a natural way, such as by gestures”. In other words, “embodied interaction is the creation, manipulation, and sharing of meaning through engaged interaction with artifacts” (Dourish 2001). Dourish adds that “embodiment is not a property of systems, technologies, or artifacts; it is a property of interaction”. Rocchesso et al. (2009) note that embodied interfaces are based on a closed loop due to motor skills, thus implying continuous and simultaneous perception and action. They argue that for embodied interfaces, the causality experienced by the users corresponds more closely to physical causality.

Wigdor and Wixon (2011) introduced the concept of natural user interfaces (NUI). “In the natural user interface, *natural* refers to the user’s behavior and feeling during the experience rather than the interface being the product of some organic process” (p. 10). They state that a device that feels truly natural is able to take full advantage of the user’s bandwidth. It may thus behave “as a sort

of appendage” (p. 10). Wigdor and Wixon provide a couple of guidelines for designing natural user interfaces (p. 13):

- Create an experience that, for expert users, can feel like an extension of their body.
- Create an experience that feels just as natural to a novice as it does to an expert user.
- Create an experience that is authentic to the medium—do not start by trying to mimic the real world or anything else.
- Build a user interface that considers context, including the right metaphors, visual indications, feedback, and input/output methods for the context.
- Avoid falling in the trap of copying existing user interface paradigms.
- Leverage innate talents and previously learned skills.

Jacob et al. (2008) reviewed the advances in post-WIMP¹ user interfaces of modern interactive systems and digital electronic devices: virtual, mixed, and augmented reality, tangible interfaces, as well as ubiquitous and pervasive computing. Many of these approaches adopt principles that are based on knowledge or skills from the *real world*, i.e. basic laws of physics or human behavior. The authors provide a theoretical framework which bases on 4 basic themes:

- *Naive Physics*: people have common sense knowledge about the physical world.
- *Body Awareness and Skills*: people have an awareness of their own physical bodies and possess skills for controlling and coordinating their bodies.
- *Environment Awareness and Skills*: people have a sense of their surroundings and possess skills for negotiating, manipulating, and navigating within their environment.
- *Social Awareness and Skills*: people are generally aware of others in their environment and have skills for interacting with them.

According to the authors, adopting these themes avoids training in interface-specific skills but rather builds on those skills that are already trained. The result may lead to reduced mental effort and thus speed up the learning process and improve the performance. Reality-based interfaces are a special case of multimodal interfaces, requiring multiple action modalities as well as multisensory feedback.

¹WIMP stands for Window, Icon, Menu, and Pointing device.

1.6. Auditory augmentation and auditory display

Stockman (2010) raised the following question: “what can we take into the research arena of auditory displays from our every day experiences of listening?”. He argued that auditory displays and interactive sonification systems should leverage the remarkable human abilities in knowledge making from sound. While the sounds that emerge from our everyday interactions reveal tons of valuable information on the involved objects themselves (more on that in Sec. 2.2), they even bear our own very characteristic signatures such as “the way we knock on doors, play instruments, type on keyboards, whistle, and even in some cases the way we breathe.” Most of the time, we aren’t even aware of the sounds we rely on in our daily routines to monitor and trigger actions. Obvious examples are the sounds of car engines, washing machines, kettles, or, not so long ago, even computers. (Stockman 2010)

Ferguson (2013) presented a framework for ambient sonification systems. Inspired by ambient information systems which have been widely investigated in the visual modality, ambient sonifications augment interactions with physical objects in a domestic environment by sound. According to the author, ambient sonifications should provide an invisible interface to ambient data representation. They should inject themselves “into a pattern of domestic behavior without significant disruption and without requiring direct attention.” The proposed interaction framework uses 5 stages which connect to a closed loop with the user: (1) interaction and associated meanings, (2) sensors, (3) data acquisition, (4) data sonification, and (5) listening and responding. Ferguson proposes the exploitation of simple in-home interactions such as turning on a light, opening a drawer or a door, or making noise, to trigger the playback of data sonifications. He further suggests that embedding sensors within physical objects themselves, instead of attaching them from the outside, may provide even better results. (Ferguson 2013)

Based on general concepts such as tangible interfaces and reality-based interaction (Sec. 1.5), (Bovermann 2009) proposed tangible auditory interfaces (TAI)—systems that combine tangible interfaces with auditory displays. Within this concept, the tangible part should provide a medium for manipulating data, algorithms, or parameters, while the aim of the auditory part is to provide the user

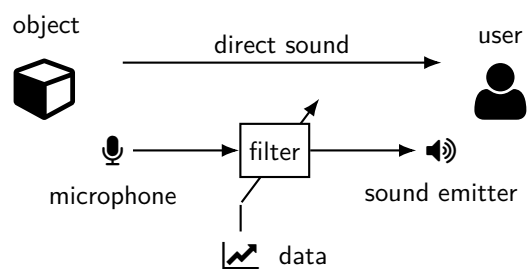


Figure 1.10.: The basic concept of auditory augmentation, adapted from Bovermann et al. (2010).

with data- and interaction-driven information. According to Bovermann et al. (2010), the key features of tangible auditory interfaces are the richness and directness of the interaction, their possibility for multi-user interaction in an ambient setting. A benefit in ergonomics is achieved by the interplay between sound and tangibility with a nature-inspired interface gestalt, so that interactions are directly derived from nature.

Within the framework of TAI, Bovermann et al. (2010) introduced auditory augmentation as a “paradigm to vary the objects’ sonic characteristics such that their original sonic response appears as augmented by an artificial sound that encodes information about external data.” Most importantly, “all this manipulation does not affect the sound’s original purpose”. This implies that a structure-borne sound is artificially altered to add an additional layer of data-inherent features. Auditory augmentation as depicted in Fig. 1.10 uses the original sound of the object as input signal of a filter which is parameterized by the given data. Note that the resulting auditory feedback is actually a sum of the original auditory feedback and the output of the filter. A successful auditory augmentation alters the sonic gestalt of the physical object, but not the presence or timing of the original auditory feedback. The ultimate goal is to display digital data as auditory features of physical objects. Bovermann et al. argue that auditory augmentation facilitates the integration of data into everyday life. However, they state that “this paradigm is neither intended nor appropriate to systematically search for specific structure in data, or even to observe exact class labels for a dataset.” The filter is usually implemented as a bank of resonators which mimic the resonances of a physical rigid object of a different material. The data is mapped to the parameters of the resonators (frequencies, decay times, and amplitudes).

1. Introduction

(Bovermann et al. 2010)

The same authors describe several case studies that aim on either data exploration or on unobtrusive data monitoring.

In *Schüttelreim*, contact microphones are attached to a hollow object filled with grainy material such as buttons or marbles. Users are thus able to explore the data by manipulating (shaking, rotating, knocking, etc.) the object, based on their knowledge gained through everyday life. The authors claim that users will learn how to shake and manipulate the interface object so that certain aspect of the data become perceptible. (Bovermann et al. 2010)

Paarreim uses two contact microphones, each attached to a small rigid object with little natural resonances, each augmented individually by a different dataset. The data is explored through physical interaction between the little objects and several kinds of materials or characteristic haptic textures. Datasets are compared by performing the same physical action with different augmented objects. Aspects of the data are explored by different surface textures. (Bovermann et al. 2010)

For *Wetterreim*, a contact microphone is attached to a physical object that people interact with on a daily basis: a computer keyboard. The filter parameters are set by real-time weather data, so that users are able to perceive differences in temperature, humidity, wind speed, or air pressure as subtle sound characteristics of the computer keyboard. Through the superposition of the filter output with the original auditory feedback, the augmented auditory feedback is perceived as one coherent auditory gestalt. In a small user study with participants testing the system for several days on their own computer, participants were able to differentiate between different weather conditions, while none of them found the system bothersome during work. However, some participants stated that they were unable to separate the data-driven part of the sound from the original auditory feedback. (Bovermann et al. 2010)

Auditory augmentation is a special case of blended sonification. The term was introduced by Tünnermann et al. (2013) as a unifying concept for “sonifications that blend into the users’ environment without confronting users with any explicitly perceived technology.” According to their working definition, “blended sonification describes the process of manipulating physical interaction sounds or environmental sounds in such a way that the resulting sound signal carries additional information

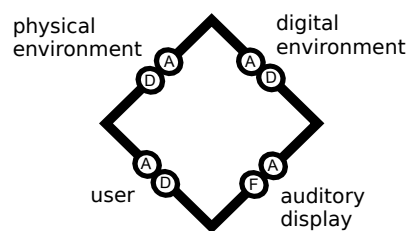


Figure 1.11.: A blank blended sonification diagram (adapted from Tünnermann et al. 2013). The filtered (F) or added (A) output of the auditory display may be driven by data (D) or audio (A) component from the user or the physical or digital environment.

of interest while the formed auditory gestalt is still perceived as coherent auditory event.” In addition, “blended sonifications should be calm, well motivated and expectable by the user. They should stay in the periphery but be ready to hand.” (Tünnermann et al. 2013)

In order to facilitate the process of sketching, comparing, and discussing blended sonifications, Tünnermann et al. (2013) developed the blended sonification diagram, as shown in Fig. 1.11. Three sides represent the three main factors: physical environment, digital environment, and user. They may contribute a data (D) and/or an audio component (A) to the auditory display which is represented by the fourth side. A connection to the added (A) output of the auditory display represents a sonification which superimposes additional sound events to the acoustic environment. A connection to the filtered (F) output represents a sonification that stays “very close to the original sound”, i.e., an auditory augmentation. In case of an interaction between multiple users, the blended sonification diagram can be expanded by additional sides for multiple users and their corresponding auditory displays. (Tünnermann et al. 2013)

Based on their practical experience, Tünnermann et al. provide guidelines that may help in the design and development process of blended sonifications. These guidelines involve 4 principles: calmness and peripheralness, coherency, expectability and familiarity, and physical origin. “Calmness connotes that sonifications should only be triggered by actions of the user and actions that take place in the ambiance of the user”. This implies that they are available at any time, if a user wants to selectively pay attention to it, but stay out of the way otherwise. Coherency refers to the recommendation that both the augmented or added sound should coherently

blend with the original auditory feedback or acoustic environment, leading to a single cognitive unit. *Expectability and familiarity* imply that the auditory feedback should be expectable and familiar as much as possible. The *physical origin* of the sonification should be perceived as being closely connected to the physical actions of the user or that of users or objects in the immediate environment of the user. (Tünnermann et al. 2013)

Weather to go by Tünnermann et al. (2014) presents an auditory weather report that informs the user about the present and upcoming weather when leaving the house. Initiated by the user opening the door, a natural soundscape is played back that represents the given weather by mimicking its natural sound (e.g., by recordings of wind, rain, and thunder).

Upstairs by Bovermann et al. (2012) is a calm communication system that aims for couples in long-distance relationships. Inspired by the footstep sounds that are commonly perceived from neighboring apartments, the footstep sounds of two spatially dislocated apartments are recorded via contact microphones, filtered, and played back in the other apartment in real time, respectively. The thereby connected persons perceive each other as living upstairs. Without much cognitive effort, a feeling of presence is created. (Tünnermann et al. 2015; Bovermann et al. 2012)

Knock'Knock by Tünnermann et al. (2013) is an auditory augmentation of a door (e.g., of an office) that provides visitors with information on the absence of the person they wish to meet. In case the person is present, he or she is able to react to a visitor knocking on the door. In absence, artificial reverberation is added to the knocking sound. The reverberation time quantifies the amount of time that has passed since the person left the office, hence unobtrusively helping the visitor to decide if it is worth to wait for the person to return, or to come back another day. (Tünnermann et al. 2013)

Slowification by Hammerschmidt and Hermann (2016) provides car drivers with real-time information on the current driving speed, relative to the currently allowed speed. The applied metaphor is that of the car radio system always traveling with the allowed speed. By spatial sound reproduction, the music or radio program is rendered to come from the back when driving way too fast, from the front when driving slower than the allowed speed, or anything in between. While the majority of the participants of a user study stated that they would prefer this system over a visual display, the results

suggest that it cannot fully replace a visual indicator. (Hammerschmidt and Hermann 2016)

Infodrops by Hammerschmidt and Hermann (2013) uses blended sonification to enhance awareness towards the consumption of limited resources in the shower. While water flow is continuously measured, the original shower sound is augmented by synthesized sounds of water drops. These are processed by tuned resonance filters which start at consonant chords and change to more and more dissonant chords as water consumption reaches 4/5 of the average, hence aiming for gradual improvement of the user's habits. (Hammerschmidt and Hermann 2013)

Bakker et al. (2010) presented two design cases for unobtrusive sonification by mechanically generated sounds.

Flunda is a coat rack that is augmented by an interactive indoor fountain. If someone hangs a coat on the rack, the corresponding tap will start to drip. As time advances, the water flow increases, hence sonifying (as well as visualizing) the time since the person arrived, with maximum water flow after 5 hours. The authors suggest that a family might benefit from this minimal information on other family members, making the question "for how long have you been home already?" redundant. (Bakker et al. 2010)

Marblelous is a physical marble run that sonifies the presence and absence of family members at home. It consists of two glass vases that hold a marble for each family member and represent home and away, respectively. If a person changes status, a marble rolls into the other vase. For continuous sonification of the current status, both vases rotate to convey information on the number of marbles by the physically-produced sound. A different texture allows to discriminate both vases by sound. (Bakker et al. 2010)

S. Barrass and T. Barrass (2013) explored the design space of embedding sonifications into physical objects. Based on the observation that sonification in outdoor activities by means of smartphone apps is a rather cumbersome and impractical, they argue that sonifications for this scope should be embedded in things, expendable and cheap to replace, and extendable and customizable. The authors presented three prototypes of such embedded sonifications.

Flotsam is a piece of wood that is hollowed out and containing a small microcontroller board that sonifies turning points of acceleration by means of impact-like clicks. The object is watertight and transmits sound by radio, to be received with radio

1. Introduction

headphones. The authors state that despite the use of headphones, the direct relationship between object and sound creates a sense of causality. (S. Barrass and T. Barrass 2013)

Jetsam implements a similar technical approach within a chunk of lava, sonifying absolute orientation via built-in mobile phone vibrators. Orientation is mapped to the frequency of a pulse train feeding the motors. The augmented object aims at guiding the orientation and position of a user's grip. (S. Barrass and T. Barrass 2013)

Lagan uses a cuttlefish backbone that is hollowed out to integrate the microcontroller and piezo-electric buzzers, so that the sound emerges directly from the object. It sonifies the continuous variation in acceleration by a sound resembling ocean waves. (S. Barrass and T. Barrass 2013)

The design studies of sonifications embedded into things that have been presented by S. Barrass and T. Barrass do not count to auditory augmentations in the strict sense of Bovermann et al. (2010), but may be regarded as blended sonifications under the premise that they maintain a certain level of calmness. They demonstrate that even very simple sonifications and limited interactions can integrate into the physical world in a useful way, if they are bound into physical objects. This points us to an aspect that is not explicitly covered by the basic concept of activity theory described in Sec. 1.4 but hidden within the category of tools: information. Digital information seems to better integrate in the physical world, if it is mediated by physical objects.

At the intersection of interaction design, sound and music computing, and auditory display, another research field has established itself. It is closely connected to the concepts of blended sonification, auditory augmentation, and interactive sonification. In *sonic interaction design (SID)*, the auditory feedback of interactive systems plays a central role, in order to take full advantage of our extraordinary sonic skills. Continuous auditory feedback is utilized to attach an additional information channel to physical objects such as tools, appliances, or artifacts, in order to enhance user experience or even extend functionality while being tightly integrated in the interaction loop. According to Franinović and Serafin (2013, p. vii), SID “explores ways in which sound can be used to convey information, meaning, and aesthetic and emotional qualities in interactive contexts. As a design practice, SID is the creative activity of shaping the relationships between artifacts, services, or environments and their users by means of interactive sound”. Several design studies

have emerged from this discipline, most of them within the category of sonifications embedded into physical objects.

Lemaitre, Houix, Franinovic, et al. (2009) developed a sonically augmented object specifically designed for studying the emotional reactions to sonic interactions. The *Flops* basically consists of a 3D-printed drinking glass that is equipped with a microcontroller that sends sensor data wireless to a computer running the sound synthesis. The sound design implements the metaphor of small virtual balls that can be poured out of the glass by tilting it. (Lemaitre, Houix, Franinovic, et al. 2009)

Lemaitre et al. (2012) evaluated the feelings that are elicited by the *Flops* in a user study. 32 different impact sounds were tested: 16 recordings of natural sounds and 16 synthetic sounds. Participants watched videos showing interaction with the *Flops* in different sound conditions. Natural sounds and sounds with low spectral centroid² were perceived as more pleasant than synthetic sounds and sounds with high spectral centroid. A second experiment examined actual interaction with the *Flops* in a game task where participants had to pour a fixed number of balls in less than 20 s. The three different sound conditions included the two sounds that induced the most negative feelings in the previous experiment, as well as the one that induced the most positive feelings. After each trial, participants had to report their feelings by means of self-assessment manikins (SAMs, Bradley and Lang 1994). In general, the range of judgments was larger than in the previous experiment. Valence was mainly influenced by the aesthetics of the sound, while arousal and dominance were mainly influenced by the functionality or difficulty level of the device. (Lemaitre et al. 2012)

Lemaitre, Houix, Visell, et al. (2009) created another artifact — the *Spinotron* — that has, similar to the *Flops*, no explicit purpose other than exploring sonic interactions. It affords only one mode of interaction, being vertical pumping of its central shaft. Integrated sensors in combination with a physical sound model create the metaphor of a ratcheted wheel whose rotation is driven by the pumping motion. The rate of rotation, audible via the frequency of impacts, increases with the energy of pumping. A first experiment showed that participants were able to estimate the speed of the rotating ratcheted wheel with fair accuracy, independent of the specific sound. A second experiment required partic-

²also called spectral center of gravity (SCG)

ipants to identify the cause of the sound by selecting among different materials, interactions, and verbal portraits. The results allowed the selection of those sound model parameters that evoked the most associations with the metaphor of a ratchet. In a third experiment, the participants' task was to keep a constant speed of the ratchet while performing with the Spinotron. Compared to a control group without auditory feedback, the auditory feedback allowed a steeper learning curve, meaning that participants actually ameliorated their performance in pumping at constant speed. (Lemaitre, Houix, Visell, et al. 2009)

A large body of research focuses on the auditory augmentation of shoes or footstep sounds (e.g., Papetti et al. 2010; Marchal et al. 2013; Lécuyer et al. 2011; Turchet et al. 2010; Batavia et al. 1997). A recent literature review is provided by Elvitigala et al. (2021). They all have in common that users wear sensor- and sometimes loudspeaker-equipped shoes to create auditory feedback that differs from the actual physical auditory feedback, in real time. The rendered auditory feedback may simulate different types of shoes (e.g., high-toe or flip-flops) or walking grounds (e.g., wood or snow), but also comprises unrealistic sound such as speech or music. The aims of these approaches range from entertainment applications to sports and medical rehabilitation. Some of these auditory augmentations will be reviewed in detail in Ch. 2 when it comes to auditory perception of physical properties.

Müller-Tomfelde and Münch (2001) presented the auditory augmentation of a pencil. Their target applications include the addition of natural auditory feedback to almost silent electronic whiteboards or graphic tablets, as well as the auditory display of windows on a computer screen by simulating different surface textures. They suggest a three-layered model for convincing auditory feedback.

The *micro surface texture* describes the finest level of surface texture. Similar to the texture of fine paper, the (tiny) peaks are homogeneous in both dimensions, without any directivity, with a stochastic spatial distribution.

The *meso surface texture* exhibits peaks that are already perceived as single sound events, including, for example, the grain of wood or sand paper. Distribution pattern might show a preferred orientation that is perceptible by sliding movement in different directions.

The *macro surface texture* is closely related to visible patterns of the surface, corresponding to indentations, bumps, or engravings, perceptible as borders.

The auditory feedback is generated separately for each of the three levels, based on physical modeling synthesis. (Müller-Tomfelde and Münch 2001)

Motor learning describes a lasting change in motor performance that is caused by training (Sigrist et al. 2013). In this context, “augmented feedback, also known as *extrinsic feedback*, is defined as information that cannot be elaborated without an external source; thus, it is provided by a trainer or a display” (Sigrist et al. 2013). *Intrinsic feedback* or internal feedback is the kind of feedback that is always present during motor learning. Sigrist et al. argue that movement sonification is very efficient for low-complexity tasks, but is assumed to add only little benefit for complex tasks. They predict, however, that error sonification might add a benefit even for complex tasks if combined with other sensory modalities.

Rocchesso et al. (2009) presented several design exercises, that are prototypical for interaction with different kitchen tools. The prototypes aim at facilitating interactions by continuous auditory feedback. A traditional mocha coffee maker was augmented with an internal force sensor to sonify the amount of pressure that is applied while screwing together the individual parts after loading with coffee and water. The sonification was inspired by the Schelleng diagram which visualizes the optimal bow force on the violin relative to bow position at constant bow velocity (Schelleng 2005). A friction sound model mimicked a glass harmonica for too loose coupling and a squeaking sound if coupled too tight. The auditory feedback was reported to be very natural, enhancing the screw connection, and making the task more engaging. The authors solved the decoupling between object and loudspeaker by mounting an actuator on the mocha itself so that the sound blends with the original auditory feedback. (Rocchesso et al. 2009)

Another design exercise described by Rocchesso et al. (2009) aims at enhancing vegetable cutting. The augmented cutting board is equipped with a contact microphone in order to detect onsets of impacts. The progress of slicing is sensed by a camera. During the task of cutting carrots, a rhythmic beat adapts to the tempo of the user.

Delle Monache et al. (2007) developed a sonically augmented dining table for investigating the closed loop between interaction, sound, and emotion. Their prototype, *the gamelunch*, is based on a prepared table which is divided into sensitive areas corresponding to the positions of specific objects on the table (dish, cutlery, glass, and decanter).

1. Introduction

Contact microphones are used to detect interactions and thus control a physically informed sound model. Augmented auditory feedback of the individual objects is rendered by loudspeakers below the table. Some objects such as salad bowl and drinking glass incorporate sensors to drive mixing and liquid pouring sounds that are congruent to the interaction. Others, however, provide feedback that contradicts the users' expectations; e.g., continuous friction sounds are played when holding the decanter. (Delle Monache et al. 2007)

The above examples show that there seem to be many different concepts which exhibit a large amount of overlap, or maybe even describe the same thing in other words. Within the context of this thesis, we stick with the term *auditory augmentation*, as it seems to be the most descriptive and self-explanatory of all the mentioned concepts, while seemingly appropriate to the case studies presented in Ch. 8. However, the original definition given by Bovermann et al. (2010), explicitly insisting on structure-borne sound, is assumed to be too strict and contradictory to the common sense of understanding of the term. In Ch. 7 we will therefore testify our working definition of auditory augmentation.

1.7. What sounds plausible?

*"I'm not concerned with plausibility;
that's the easiest part of it, so why bother?"*
— Alfred Hitchcock (Truffaut et al. 1984, p. 99)

In this section, we will summarize approaches and definitions concerning plausibility that might be relevant within the context of HCI and psychophysics.

Connell and Keane (2006) found that the term plausibility had never been well explained in previous literature, although we humans judge the plausibility of information on a daily basis, and although plausibility has been studied in various different contexts of research. For example, plausibility judgment as a cognitive strategy avoids information retrieval from long-term memory and speeds up cognitive processes (Reder 1982). Connell and Keane complain that despite a solid body of research on the influence of plausibility on text comprehension, reasoning, or the solution of arithmetic problems, plausibility is typically treated as a rather unspecific variable that is interrogated without further explanation. Nevertheless, they found a rough consensus that plausibility judgment involves some kind of

concept-coherence, i.e., how a certain scenario conceptually coheres with prior knowledge. According to Connell and Keane, "the plausibility of category membership can be viewed as a function of how well the object's features cohere with one another, according to prior knowledge of causal relations between category features".

Connell and Keane state that judging the plausibility of event scenarios requires two steps. First, finding causal chains based on prior knowledge in order to make the necessary inferences. Second, evaluating the match between these inferences and what has been experienced in the past. A simple example sentence: "The balloon landed on the pin and melted." Even if both actions are plausible in isolation, the scenario is most likely judged as rather implausible, because an inferred connection would involve a high amount of conjecture about the possible heat of the pin making the balloon melt. (Connell and Keane 2006)

Connell and Keane derived a cognitive model of plausibility, in an attempt to predict human plausibility judgment. The basis of the model—their so-called knowledge-fitting theory— involves two processing stages. In the *comprehension stage* a mental representation of a given scenario is created from verbal description and inferences based on prior knowledge. In the *assessment stage* the mental representation is compared to prior knowledge, i.e., its concept-coherence. Within the knowledge-fitting theory of Connell and Keane, a plausible scenario is one that fits prior knowledge

1. *using many different sources of corroboration.* The scenario should have several distinct pieces of prior knowledge supporting any necessary inferences.
2. *without complex explanation.* The scenario must be represented without relying on extended or convoluted justifications.
3. *using minimal conjecture.* The scenario must be represented by avoiding the introduction of hypothetical entities wherever possible (i.e., no deus ex machina). (Connell and Keane 2006)

Equation 1.1 summarizes the plausibility model in terms of a theoretical function:

$$\begin{aligned} \text{plausibility} &= 1 - \text{implausibility} \\ \text{implausibility} &= \frac{\text{complexity}}{\text{corroboration} - \text{conjecture}} \end{aligned} \quad (1.1)$$

Hence, "a scenario will be perfectly plausible only if its representation has minimal *complexity* and

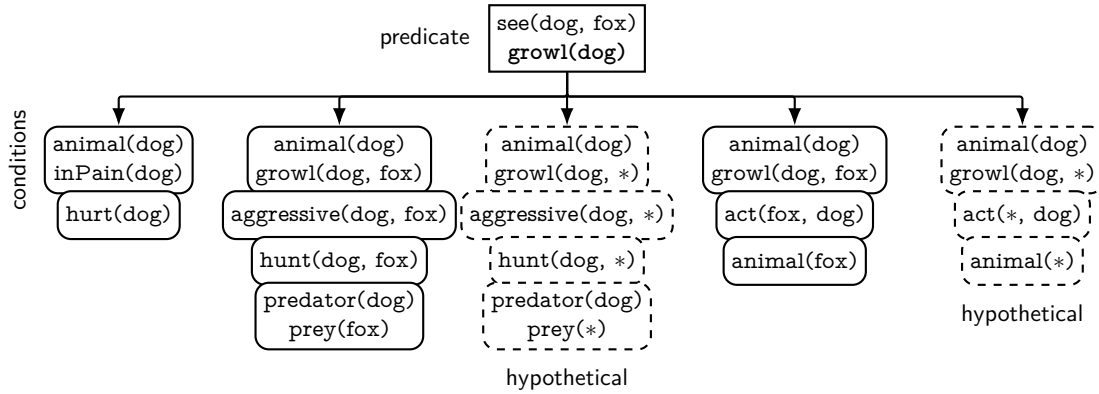


Figure 1.12.: Example of a formalized scenario tree created within the comprehension stage of PAM for the scenario “The pack saw the fox. The hounds growled.” The wildcard character * represents hypothetical information that is not given by the scenario.

conjecture, and/or maximal corroboration” (Connell and Keane 2006).

A computational implementation of the knowledge-fitting theory is provided by Connell and Keane through the *plausibility analysis model (PAM)*. It comprises two stages.

In the *comprehension stage*, a scenario is converted to a formalized conditional tree of individual predicates and chains of conditions. As an example, we analyze the following scenario: “The pack saw the fox. The hounds growled.” The formalized scenario tree is depicted in Fig. 1.12. It comprises 5 paths of which two incorporate hypothetical knowledge. Not all paths make use of all the given information. In the *assessment stage*, a plausibility rating is computed from the scenario tree, based on the average path length L (representing the complexity of explanation), the total number of paths P (representing corroboration), and the number of non-hypothetical paths N (representing explanations without conjecture). Equation 1.2 then gives the implausibility $\bar{\Pi}$:

$$\bar{\Pi} = \frac{L}{L + 1} \frac{N}{P + \frac{N}{P}} \quad (1.2)$$

Implausibility $\bar{\Pi}$ (ranging from 0 to 1) is converted to plausibility Π in analogy to the simplified Eq. 1.1, but with additional squaring:

$$\Pi = (1 - \bar{\Pi})^2 \quad (1.3)$$

Figure 1.13 depicts how the model plausibility depends on L , P , and N . The example from Fig. 1.12

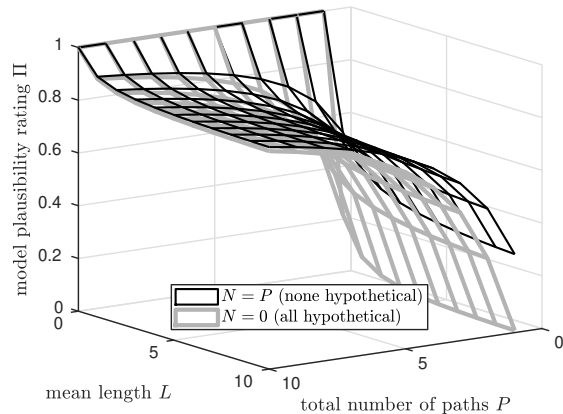


Figure 1.13.: Model plausibility rating Π depending on the total number of paths P , their average length L , and the amount of non-hypothetical paths N (adapted from Connell and Keane 2006).

includes $P=5$ total paths of average length $L=3.2$, of which $N=3$ are non-hypothetical. This leads to a plausibility rating of $\Pi=0.75$. (Connell and Keane 2006)

PAM was evaluated on scenarios constructed from two short descriptive sentences. For each scenario, participants rated plausibility between 0 (implausible) and 10 (plausible); we divide the values by 10 for to ensure a range between 0 and 1. The model was able to successfully predict human plausibility ratings ($R^2=0.603$). Furthermore, different types of inferences (causal, attributional, temporal, unrelated) led to significantly different human plausibility ratings. Average values were 0.83 (causal), 0.61 (attributional), 0.55 (temporal), and 0.15 (un-

1. Introduction

related). Even this order pattern was reproduced by the model. (Connell and Keane 2006)

Lindau and Weinzierl (2012) evaluated auditory virtual environments with respect to their perceptual plausibility. They define a plausible virtual environment as “a simulation in agreement with the listener’s expectation towards a corresponding real event”. These expectations do not refer to an exact physical or perceptual identity of reality or simulation, but rather to an inner reference resulting from personal experience and expectations. Consistent with other authors, they note that plausibility is highly dependent of the actual application and task. For example, in the context of game audio, a rather pragmatic concept of plausibility is prevailing: “as long as there is no obvious contradiction between the visual and the acoustic representation of a virtual scene, the human senses merge auditory and visual impressions” (Grimshaw 2011, p. 159).

For the experimental evaluation of plausibility, Lindau and Weinzierl (2012) argue against an absolute rating, as this would lead to a strong individually different response bias due to “personal theories about the credibility of virtual realities and the performance of media systems in general”. They instead suggest a criterion-free assessment of plausibility by using a yes/no paradigm where participants simply decide between simulation (yes) and recording (no). This yields an evaluation with regard to an (unknown!) inner reference, in contrast to an external, given reference. By applying standard methods of signal detection theory (Macmillan and Creelman 2005), the implausibility can then be expressed as the sensitivity d' which is computed from the probability of correct responses P_c :

$$d' = \sqrt{2}\Phi(P_c) \quad (1.4)$$

with $\Phi(p)$ being the inverse cumulative normal distribution. A simulation is assumed to be plausible if P_c is less than 55 %, meaning that it exceeds pure guessing by less than 5 %. (Lindau and Weinzierl 2012)

Coco and Duran (2016) examined cross-modal interactions between plausibility and sensory congruency. In their experiment, each trial consisted of a textual stimulus in form of a descriptive sentence, and a visual stimulus in form of a photo composition. The pairs of stimuli were presented in different combinations of two types of congruency (congruent, incongruent) and two types of plausibility (plausible, implausible). Congruency indicated whether content was matched *between* stimuli, while plausibility indicated whether the content was matched

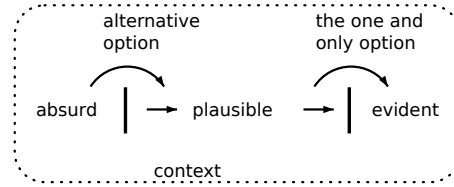


Figure 1.14.: Plausibility model (adapted from Böhnert and Reszke 2014).

within stimuli. For example, the sentence could be “the boy is eating a hamburger” (plausible) or “the boy is eating a brick” (implausible), whereas the equivalent visual stimulus was a photo composition of a boy eating a hamburger or a brick, respectively. For 900 pairs of textual and visual stimuli (both presented only for a limited time), participants indicated if both stimuli were congruent (yes/no) and rated plausibility, visual saliency of the target object, congruency between scene and sentence, and grammaticality of the sentence. Overall, participants achieved high accuracy (87 %) in the identification of congruency. Answers were significantly more accurate for plausible stimuli than for implausible ones. In case of implausibility, accuracy was significantly lower for congruent stimuli. An analysis of response times and mouse movements revealed an early hesitation and bias to identify implausible stimuli as incongruent. The authors infer that the plausibility of stimuli mediates congruency expectations in a verification task. They conclude that congruency between different sources of information alone is not sufficient for facilitating cognitive processing. The benefit instead depends strongly on the plausibility of the processed information. (Coco and Duran 2016)

Böhnert and Reszke 2014 analyzed the usage of the term “plausibility” in scientific writing. They observed that it is frequently used as a matter of course, while its meaning is almost never explicitly discussed or explained. Based on its usage and context found in the literature, Böhnert and Reszke derived a model of plausibility that characterizes the evolution of scientific theories, but might also be useful for understanding plausibility judgment in general. According to the authors, the beginning in the formation of scientific facts is a theory that is not anymore regarded as absurd (see Fig. 1.14). If a majority of researchers consider it as an alternative option, it reaches the threshold of plausibility. The classification as being plausible implies that there are other (likewise plausible) alternatives to it. If

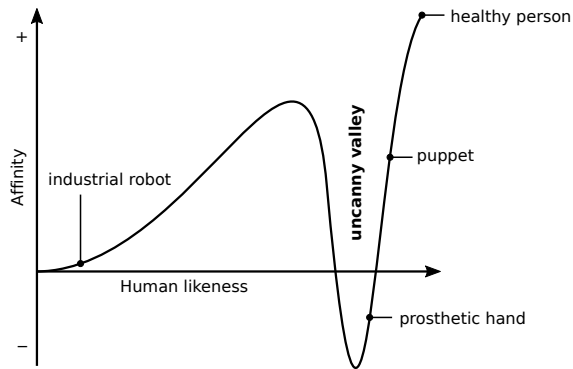


Figure 1.15.: The uncanny valley. Adapted from Mori et al. (2012).

a plausible theory manages to preclude all other alternatives, it becomes the one and only evident option, and the matter it describes becomes an obvious matter of fact. This model is valid only within a given context or comprehension environment of the scientific discourse. (Böhnert and Reszke 2014)

Plausibility is sometimes used as a synonym for realism. What does realism mean and in what aspects does it differ from plausibility?

Wages et al. (2004) examined the usage of the term “realism” with regard to computer games and inferred that it is often misconceived in this context. Against the common belief that more realism equals higher believability, they even suggest that realism works counterproductive for the immersion of the user. They argue that simply copying nature neglects the fact that from all the incoming information, a vast majority is considered as non-essential by human senses and filtered out physiologically or mentally. Studies on animal behavior, for example, showed that a simple strip of cardboard—although far from realistic—may induce greater social interaction than a stuffed animal. Wages et al. argue that also for humans, the perceptual discrepancies compared to reality are increasingly weighted by our cognition, the closer an artificial stimulus approaches reality. An increase in realism might therefore paradoxically lead to decreased believability. This dip in the affinity or empathy curve, that appears with increasing level of realism, just before a robot (or similar character) gets perceptually indistinguishable from reality, is called the uncanny valley (see Fig. 1.15). While attempts have been made to relate the uncanny valley to sound (e.g., Grimshaw 2009), it is unlikely that something like an audio uncanny valley exists.

Wages et al. (2004) give another paradox of re-

alism: stimuli interdependencies. In a virtual environment, a stimulus is actually a composition of a multitude of different stimuli. With a certain aspect of the complex stimulus advancing towards perfection, the shortcomings of other aspects might be revealed, leading to a paradoxical growth in a spectator’s disbelief. The authors cite the work of Hodgins et al. (1998) who demonstrated that an improvement of a virtual character’s visual presentation does not automatically lead to an increase of its realism. Wages et al. argue that “the increase of one aspect of realism (visual presentation) can rapidly lead to much more weight of another aspect (animations)”. Flaws in one aspect might therefore not be noticed until another aspect gets optimized.

Closely connected to realism is the concept of presence, as introduced by Lombard and Ditton (1997). The authors interpret presence as a perceptual illusion that is a property of a person. They constitute 6 interrelated but distinct conceptualizations of presence throughout the literature. If interpreted as *social richness*, “presence is the extent to which a medium is perceived as sociable, warm, sensitive, personal or intimate when it is used to interact with other people”. Presence may also describe *realism*, i.e., “the degree to which a medium can produce seemingly accurate representations of objects, events, and people”. A definition in terms of *transportation* might relate it to a user or object being transported to another place. Presence is often used as a synonym for *immersion*, i.e., “the degree to which a virtual environment submerges the perceptual system of the user” (Biocca and Levy 1995, p. 57). The final two conceptualizations refer presence to a kind of *social actor* or *medium* concerning the relationship and interaction between users.

In summary, we identify at least two entirely different concepts of plausibility. The first seems to be strongly related to the congruency between different sensory channels. A multisensory stimulus remains plausible, as long as no obvious contradiction occurs between the information from different sensory channels. This type of plausibility is usually evaluated unconsciously. We become aware of it only if we either consciously reflect about it, or if incongruent information is perceived. We will further call this low-level plausibility. The second type of plausibility seems to be evaluated afterwards, when the perceived stimulus is interpreted consciously. The interpretation may include a physical explanation of the stimulus. For example, the auditory feedback of a small solid object may feel completely plausible at

first, but may get implausible after thinking about it, if the sound is too low-pitched for being physically feasible. This conscious concept we call high-level plausibility.

1.8. Conclusions

Auditory feedback is the sound we perceive in reaction to our physical interactions. As it conveys information about the ongoing physical processes, it helps us to adapt our unconscious operations to the given conditions and thus reach the goals of our conscious actions. The thus created information channel lets us stay within the sonic interaction loop. More generally, the interaction loop, as achieved through the integration of all available sensory modalities, allows that the tools we use to mediate our actions effectively become an extension of our body. With growing complexity of these tools, e.g., from basic hand tools via mechanic machines to digital devices, the amount of useful information contained in the original auditory feedback decreases. This lack of usable auditory feedback in modern tools highlights the need for sonic interaction design in general and auditory augmentation in particular. In its strict sense, the original auditory feedback is modulated in order to become augmented auditory feedback: to facilitate the perception of the encoded information, to change the encoded information, or to add new information.

We assume that to become accepted by the users, augmented auditory feedback needs to be plausible and usable with respect to the given physical interaction, and with respect to information from other sensory modalities. These requirements limit the parameter space for auditory augmentation and thus limit the total information capacity of the associated communication channel to the manifold of alternative auditory feedbacks.

While usability is generally well-defined, the meaning of plausibility varies across the literature and lacks a theoretical model for non-speech information. We identify two levels of plausibility: low-level plausibility and high-level plausibility. Both have in common that the perceived information is compared to previously acquired knowledge. Low-level plausibility describes the congruency between the information from different sensory modalities (audition, vision, touch, etc.). A stimulus can, however, already be plausible, if there is no contradiction between the involved information channels. High-level plausibility involves reasoning and the derivation of

meaning from the gathered information, and therefore doesn't affect the stimulus itself but rather its interpretation: a scenario or action that may have produced the specific stimulus. High-level plausibility is well explained by the plausibility analysis model (PAM) by Connell and Keane (2006).

Especially interactions with the digital environment benefit from reality-based interactions that make digital information tangible and thus allow users to behave naturally by projecting their knowledge from the physical environment. Building on top of established concepts within sonification and sonic interaction design, auditory augmentation in the broader sense may also add additional sound to a physical object to convey information.

In Ch. 2 we will now review what kinds of physical information human listeners are able to extract from auditory feedback.

Bibliography

- Bakker, Saskia et al. (2010). "Sounds like home: Sonification and physical interaction in the periphery and center of attention." In: *Interactive Sonification Workshop (ISon)*, pp. 55–58.
- Barrass, Stephen and Tim Barrass (2013). "Embedding Sonifications in Things". In: *International Conference on Auditory Display (ICAD)*. Łódź, Poland, pp. 149–152.
- Batavia, Mitchell, John G. Gianutsos, and Markella Kambouris (Dec. 1997). "An augmented auditory feedback device". In: *Archives of Physical Medicine and Rehabilitation* 78.12, pp. 1389–1392. DOI: 10.1016/S0003-9993(97)90317-8.
- Biocca, Frank and Mark R. Levy, eds. (Mar. 1, 1995). *Communication in the Age of Virtual Reality*. New York: Routledge. DOI: 10.4324/9781410603128.
- Böhnert, Martin und Paul Reszke (2014). „Linguistisch-philosophische Untersuchungen zu Plausibilität: Über kommunikative Grundmuster bei der Entstehung von wissenschaftlichen Tatsachen". In: *Proceedings der 1. Tagung des Nachwuchsnetzwerks "INSIST"*, S. 40–67.
- Bovermann, Till (2009). "Tangible auditory interfaces: combining auditory displays and tangible interfaces." PhD thesis. Universität Bielefeld.
- Bovermann, Till, René Tünnermann, and Thomas Hermann (Apr. 2010). "Auditory Augmentation". In: *International Journal of Ambient Computing and Intelligence* 2.2, pp. 27–41. DOI: 10.4018/jaci.2010040102.

- Bovermann, Till et al. (2012). "Upstairs – Supporting Peripheral Awareness Between Non-Colocated Spaces". In: International Conference on Pervasive Computing.
- Bradley, Margaret M. and Peter J. Lang (Mar. 1994). "Measuring emotion: The self-assessment manikin and the semantic differential". In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1, pp. 49–59. DOI: 10.1016/0005-7916(94)90063-9.
- Castiello, Umberto et al. (Aug. 17, 2010). "When Ears Drive Hands: The Influence of Contact Sound on Reaching to Grasp". In: *PLoS ONE* 5.8, e12240. DOI: 10.1371/journal.pone.0012240.
- Coco, Moreno I. and Nicholas D. Duran (Dec. 2016). "When expectancies collide: Action dynamics reveal the interaction between stimulus plausibility and congruency". In: *Psychonomic Bulletin & Review* 23.6, pp. 1920–1931. DOI: 10.3758/s13423-016-1033-6.
- Connell, Louise and Mark T. Keane (Jan. 2, 2006). "A Model of Plausibility". In: *Cognitive Science* 30.1, pp. 95–120. DOI: 10.1207/s15516709cog0000_53.
- Delle Monache, Stefano et al. (2007). "Gamelunch: a Physics-Based Sonic Dining Table." In: *International Computer Music Conference (ICMC)*.
- Dourish, Paul (2001). *Where the action is: the foundations of embodied interaction*. Cambridge, Mass: MIT Press. ISBN: 978-0-262-04196-6.
- Elvitigala, Don Samitha, Jochen Huber, and Suranga Nanayakkara (Feb. 22, 2021). "Augmented Foot: A Comprehensive Survey of Augmented Foot Interfaces". In: *Augmented Humans Conference 2021*. Rovaniemi Finland: ACM, pp. 228–239. DOI: 10.1145/3458709.3458958.
- Ferguson, Sam (2013). "Sonifying every day: activating everyday interactions for ambient sonification systems". In: *International Conference on Auditory Display (ICAD)*. Łódź, Poland, pp. 77–84.
- Franinović, Karmen and Stefania Serafin, eds. (2013). *Sonic interaction design*. Cambridge, Massachusetts: The MIT Press.
- Furfaro, Enrico et al. (2013). "Sonification of surface tapping: influences on behavior, emotion and surface perception." In: *Interactive Sonification Workshop (ISon)*. Erlangen, Germany.
- Furfaro, Enrico et al. (2015). "Sonification of virtual and real surface tapping: evaluation of behavior changes, surface perception and emotional indices". In: *IEEE MultiMedia*, pp. 1–1. DOI: 10.1109/MMUL.2015.30.
- Gerrig, Richard J. (2013). *Psychology and life*. 20th ed. Boston: Pearson. ISBN: 978-0-205-85913-9.
- Grimshaw, Mark (2009). "The audio Uncanny Valley: Sound, fear and the horror game." In: *Games Computing and Creative Technologies*.
- ed. (2011). *Game sound technology and player interaction: concepts and development*. Hershey PA: Information Science Reference.
- Groß-Vogt, Katharina (Sept. 15, 2020). "The drinking reminder: prototype of a smart jar". In: *Proceedings of the 15th International Conference on Audio Mostly*. Graz Austria: ACM, pp. 257–260. DOI: 10.1145/3411109.3411130.
- Großhauser, Tobias and Thomas Hermann (2010). "Multimodal closed-loop human machine interaction". In: *Interactive Sonification Workshop (ISon)*. Stockholm, Sweden.
- Hammerschmidt, Jan and Thomas Hermann (2013). "Infodrops: Sonification for enhanced awareness of resource consumption in the shower". In: *International Conference on Auditory Display (ICAD)*. Łódź, Poland, pp. 57–64.
- (2016). "Slowification: An in-vehicle auditory display providing speed guidance through spatial panning." In: *Interactive Sonification Workshop (ISon)*. Bielefeld, Germany.
- Hartson, H. Rex and Pardha S. Pyla (2012). *The UX Book: process and guidelines for ensuring a quality user experience*. Amsterdam; Boston: Elsevier. ISBN: 978-0-12-385241-0.
- Hodgins, J.K., J.F. O'Brien, and J. Tumblin (Dec. 1998). "Perception of human motion with different geometric models". In: *IEEE Transactions on Visualization and Computer Graphics* 4.4, pp. 307–316. DOI: 10.1109/2945.765325.
- ISO (2009). *9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems*. International Organization for Standardization.
- Jacko, Julie A (2012). *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. 3rd ed. Boca Raton, Fla: CRC, Taylor & Francis. ISBN: 978-1-4398-2943-1.
- Jacob, Robert J.K. et al. (2008). "Reality-based interaction: a framework for post-WIMP interfaces". In: *CHI conference on human factors in computing systems*. Florence, Italy: ACM. DOI: 10.1145/1357054.1357089.
- Kaptelinin, Victor (1995). "Activity Theory: Implications for Human-Computer Interaction". In: *Context and Consciousness: Activity Theory and*

1. Introduction

- Human-Computer Interaction*. Massachusetts Institute of Technology, pp. 103–116.
- Kuutti, Kari (1995). “Activity Theory as a Potential Framework for Human-Computer Interaction Research”. In: *Context and Consciousness*. Ed. by Bonnie A. Nardi. The MIT Press. DOI: 10.7551/mitpress/2137.003.0006.
- Lécuyer, Anatole et al. (2011). “Shoes-Your-Style: Changing Sound of Footsteps to Create New Walking Experiences”. In: *Workshop on Sound and Music Computing for Human-Computer Interaction (CHIItaly)*.
- Lemaitre, Guillaume, Olivier Houix, K Franinovic, et al. (2009). “The flops glass: a device to study emotional reactions arising from sonic interactions”. In: *Sound and Music Computing Conference (SMC)*.
- Lemaitre, Guillaume, Olivier Houix, Yon Visell, et al. (Nov. 2009). “Toward the design and evaluation of continuous sound in tangible interfaces: The Spinotron”. In: *International Journal of Human-Computer Studies* 67.11, pp. 976–993. DOI: 10.1016/j.ijhcs.2009.07.002.
- Lemaitre, Guillaume et al. (July 2012). “Feelings Elicited by Auditory Feedback from a Computationally Augmented Artifact: The Flops”. In: *IEEE Transactions on Affective Computing* 3.3, pp. 335–348. DOI: 10.1109/T-AFFC.2012.1.
- Lindau, Alexander and Stefan Weinzierl (Sept. 1, 2012). “Assessing the Plausibility of Virtual Acoustic Environments”. In: *Acta Acustica united with Acustica* 98.5, pp. 804–810. DOI: 10.3813/AAA.918562.
- Lombard, Matthew and Theresa Ditton (Sept. 1, 1997). “At the Heart of It All: The Concept of Presence.” In: *Journal of Computer-Mediated Communication* 3.2. DOI: 10.1111/j.1083-6101.1997.tb00072.x.
- Ma, Zhaoyuan et al. (June 2015). “Haptic keyclick feedback improves typing speed and reduces typing errors on a flat keyboard”. In: *World Haptics Conference (WHC)*. Evanston, IL: IEEE, pp. 220–227. DOI: 10.1109/WHC.2015.7177717.
- Macmillan, Neil A. and C. Douglas Creelman (2005). *Detection theory: a user’s guide*. 2nd ed. Mahwah, N.J: Lawrence Erlbaum Associates. ISBN: 978-0-8058-4230-2.
- Marchal, Maud et al. (2013). “Multimodal Rendering of Walking Over Virtual Grounds”. In: *Human Walking in Virtual Environments*. Ed. by Frank Steinicke et al. New York, NY: Springer New York, pp. 263–295. DOI: 10.1007/978-1-4419-8432-6_12.
- Markosian, Ned (Sept. 2000). “What are Physical Objects?” In: *Philosophy and Phenomenological Research* 61.2. DOI: 10.2307/2653656.
- Martín, Rodrigo et al. (2015). “Multimodal perception of material properties”. In: *ACM SIGGRAPH Symposium on Applied Perception*. Tübingen, Germany: ACM Press, pp. 33–40. DOI: 10.1145/2804408.2804420.
- Mori, Masahiro, Karl MacDorman, and Norri Kageki (June 2012). “The Uncanny Valley”. In: *IEEE Robotics & Automation Magazine* 19.2, pp. 98–100. DOI: 10.1109/MRA.2012.2192811.
- Müller-Tomfelde, Christian and Tobias Münch (2001). “Modeling And Sonifying Pen Strokes On Surfaces”. In: *Conference on Digital Audio Effects (DAFX)*. Limerick, Ireland.
- Nigay, Laurence and Joëlle Coutaz (1993). “A Design Space For Multimodal Systems: Concurrent Processing and Data Fusion”. In: *Proceedings of the INTERACT’ and CHI conference on Human factors in computing system*, pp. 172–178.
- Norman, Donald A. (2013). *The design of everyday things*. Revised and expanded edition. New York, New York: Basic Books. ISBN: 978-0-465-05065-9.
- Papetti, Stefano et al. (2010). “Audio-tactile Display of Ground Properties Using Interactive Shoes”. In: *International Workshop on Haptic and Audio Interaction Design*. Ed. by Rolf Nordahl et al. Red. by John C. Mitchell et al. Springer Berlin Heidelberg, pp. 117–128. DOI: 10.1007/978-3-642-15841-4_13.
- Reder, Lynne M. (1982). “Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification.” In: *Psychological Review* 89.3, pp. 250–280. DOI: 10.1037/0033-295X.89.3.250.
- Rocchesso, Davide, Pietro Polotti, and Stefano Delle Monache (2009). “Designing Continuous Sonic Interaction”. In: *International Journal of Design* 3.3, pp. 13–25.
- Schelleng, John C. (Aug. 12, 2005). “The bowed string and the player”. In: *The Journal of the Acoustical Society of America* 53.1. DOI: 10.1121/1.1913322.
- Sedda, Anna et al. (Mar. 2011). “Integration of visual and auditory information for hand actions: preliminary evidence for the contribution of natural sounds to grasping”. In: *Experimental Brain Research* 209.3, pp. 365–374. DOI: 10.1007/s00221-011-2559-5.
- Serafin, Stefania et al. (2007). “Audio-Haptic Physically Based Simulation and Perception of Contact

- Textures.” In: *International Conference on Auditory Display (ICAD)*. Montréal, Canada, pp. 203–207.
- Sigrist, Roland et al. (Feb. 2013). “Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review”. In: *Psychonomic Bulletin & Review* 20.1, pp. 21–53. DOI: 10.3758/s13423-012-0333-8.
- Stockman, Tony (2010). “Listening to People, Objects, and Interactions”. In: *Interactive Sonification Workshop (ISon)*. Stockholm, Sweden, pp. 3–8.
- Truffaut, François, Alfred Hitchcock, and Helen G. Scott (1984). *Hitchcock*. Rev. ed. New York: Simon and Schuster. ISBN: 978-0-671-52601-6.
- Tünnermann, René, Jan Hammerschmidt, and Thomas Hermann (2013). “Blended sonification—sonification for casual information interaction.” In: *International Conference on Auditory Display (ICAD)*. Lodz, Poland.
- Tünnermann, René et al. (2014). “Weather to go – a blended sonification application”. In: *International Conference on Auditory Display (ICAD)*. New York, USA.
- Tünnermann, René et al. (2015). “Upstairs: A calm auditory communication and presence system.” In: *International Conference on Auditory Display (ICAD)*. Graz, Austria.
- Turchet, Luca et al. (Oct. 2010). “Audio-haptic physically-based simulation of walking on different grounds”. In: *IEEE Multimedia Signal Processing Conference (MMSP)*. Saint-Malo, France: IEEE, pp. 269–273. DOI: 10.1109/MMSP.2010.5662031.
- Turk, Matthew (Jan. 2014). “Multimodal interaction: A review”. In: *Pattern Recognition Letters* 36, pp. 189–195. DOI: 10.1016/j.patrec.2013.07.003.
- Ullmer, B. and H. Ishii (2000). “Emerging frameworks for tangible user interfaces”. In: *IBM Systems Journal* 39.3, pp. 915–931. DOI: 10.1147/sj.393.0915.
- VandenBos, Gary R., ed. (2015). *APA dictionary of psychology*. 2nd ed. Washington, DC: American Psychological Association.
- Wages, Richard, Stefan M. Grünvogel, and Benno Grützmacher (2004). “How Realistic is Realism? Considerations on the Aesthetics of Computer Games”. In: *Entertainment Computing (ICEC)*. Ed. by M. Rauterberg. Vol. 3166. Springer Heidelberg, pp. 216–225. DOI: 10.1007/978-3-540-28643-1_28.
- Wigdor, Daniel and Dennis Wixon (2011). *Brave NUI world: designing natural user interfaces for touch and gesture*. Burlington, Mass: Morgan Kaufmann. ISBN: 978-0-12-382231-4.

2. The psychophysics of auditory feedback: an annotated bibliography

We humans are already experts when it comes to the exploitation of sounds for knowledge making—especially if they emerge from physical processes. For about any sound we readily have some idea on its physical origin. Especially if the underlying physical process involves a living being, we are generally very sensitive to the information that is encoded in the sound. However, also non-living physical objects are usually identified with great detail, without even looking or touching them, just by using our ears. We all, not only blind people, make use of the auditory cues while navigating within the house or apartment, or even on the streets in case of low light—even if we are not all the time aware of this fact. This chapter is a literature review—maybe more a vast aggregation of literature—concerning the human capabilities in the identification of physical sounds and their underlying physical parameters. As sound per se rarely comes alone, not only the unisensory (auditory) case is considered, but also multisensory perception that involves auditory stimuli in conjunction with haptic and/or visual stimuli. The references are structured on the basis of physical properties.

This chapter is organized as follows. First, in Sec. 2.1, we examine sound source identification from a general perspective or level of information (“who or what does what with what?”). We summarize general research on the identification of inanimate and animate sounds, as well as the effects of causal uncertainty and informational masking. In Sec. 2.2 we take a detailed look at the auditory perception of every single physical parameter that might be relevant to auditory augmentation. In Sec. 2.3, the physiological limitations in the perception of the sound parameters that convey the information on physical parameters are summarized. Section 2.4 contains a brief summary of approaches for robotic perception of physical parameters. In Sec. 2.5, literature on multisensory perception of physical parameters as well as on the interplay between the different sensory modalities with auditory feedback is reviewed. In contrast to the other chapters of this thesis, references of this chapter are

provided within the individual (sub-)sections.

2.1. Source identification, causal uncertainty, and informational masking

Let us begin at the start of it all: the moment at which a certain sound event is triggered, e.g., through a physical action of a living being. How long does it take to identify it?

Guillaume et al. (2004) measured the time it takes to identify the physical cause of a sound. 117 pre-recorded sounds were presented to participants performing a free identification task. Average times reached from 50 ms (*dog bark*) to 675 ms (*teeth brushing*). 34 sounds were identified within 100 ms. Some selected sounds originated from impacted rigid objects and supposedly include material and/or size and shape identification: *fork knocking a plate* (100 ms), *coin* (108 ms), *xylophone* (120 ms), *broken glass* (133 ms and 179 ms), *beaten violin* (143 ms), *pans* (170 ms), *spoon knocking a plate* 270 ms, *ping pong ball* 425 ms.

Giordano and McDonnell (2007) examined the relevance of acoustical and conceptual information to the perceived similarity of sound events (see also Giordano et al. 2010). In a free identification task, participants labeled a large set of recorded animate and inanimate environmental sounds by a verb and up to two nouns. The authors found a significantly higher naming agreement for animate sounds than for inanimate sounds, suggesting a larger vocabulary for objects than for actions. In a following hierarchical sorting experiment, participants estimated the similarity between pairs of sounds in 3 conditions: based on acoustical information, based on conceptual information, and unbiased without specification of criteria. In the acoustical condition, participants were able to effectively suppress contextual information on similarity ratings. The same applies in the conceptual condition the other way around. However, both types of judgments were

not completely independent. For unbiased similarity ratings, participants exploited both acoustical as well as contextual information. The authors suggest that in everyday life, we generally hear animate concepts and inanimate sound sources. (Giordano and McDonnell 2007)

It is possible that two entirely different physical processes lead to the same sound or at least cannot be distinguished by our ears. This leads to causal uncertainty in the identification of sounds: the more alternative causes for a specific sound exist, the higher its causal uncertainty.

Ballas performed a large number of experiments to investigate the causal uncertainty when identifying everyday sounds (e.g., Ballas 1993). He proposed to derive the causal uncertainty based on the variation of responses gathered in a free identification task. The procedure involves the sorting of identification responses into categories of equivalent events or synonyms. This sorting is performed by humans based on objective criteria. The apparent problem of differences between sorters is solved by using multiple sorters and evaluating the level of agreement by cross-correlation analysis. The sorted responses form the basis for computing the causal uncertainty H_{cu} via the joint entropy between the J different objects and K different actions identified for a certain sound i :

$$H_{cui} = - \sum_j^J \sum_k^K p_{ijk} \log_2(p_{ijk}) \quad (2.1)$$

where p_{ijk} is the proportion of verbalizations that was used for describing an object j and action k . According to Ballas (1993), causal uncertainty correlates strongly with identification time and accuracy, and weakly with ecological frequency. In case of impact sounds, however, causal uncertainty was higher than what an estimation based on ecological frequency would predict. Sounds exhibiting a high typicality led to faster response times than atypical sounds. Furthermore, the identification of sounds with alternative causes and thus their causal uncertainty was influenced by the particular context.

A comprehensive study on the identification and categorization of environmental sounds was provided by Lemaitre et al. (2010). Environmental sounds in this context are regarded as audible acoustic events that are caused by motions in the everyday human environment. According to the authors, such sounds can be categorized according to three types of similarities: acoustical similarities according to acoustical properties, causal similarities according

to the identified physical event causing the sound, and semantic similarities according to some other kind of knowledge or meaning interpreted by the listeners. In a first experiment, 96 recorded sounds were selected that usually occur in a kitchen. For each sound, participants freely identified the object by a noun and the action by a verb. The gathered responses were then sorted into categories of similar events by different persons based on formal criteria, adopting the procedure proposed by Ballas (1993); a correlation between the sorters' results showed a high reliability. Subsequently, the causal uncertainty H_{cu} was computed to be between 0 and 4.61 (median values for the individual sounds). 60 of these sounds were randomly selected as stimuli for a second experiment. In this case, participants were either experts (professional musicians, sound engineers, etc.) or non-experts. Their task was to sort the stimuli into an arbitrary number of categories by using a graphical user interface. For each category, properties had to be entered that are shared by the category members. Participants were then asked which of the 3 criteria they had used: acoustical, causal, or semantic similarities, or optionally some other or unknown criteria. On average, participants created 11.5 categories. 32.2% were grouped by acoustical, 45% by causal, and 12.5% by semantic similarities. Experts decided more on the basis of acoustical, non-experts on causal similarities. The authors infer, in accordance with previous studies, that judging the sound based on acoustical properties requires prior training, either implicitly or explicitly. Sounds with high causal uncertainty were often grouped together according to acoustical similarities; at low causal uncertainty, non-experts switched to grouping by causal similarities. In a follow-up experiment, only a moderate correlation was found between confidence and causal uncertainty ($r = -0.58$). Nevertheless, (inverse) confidence allowed to draw the same conclusions on the previous experiment as causal uncertainty. Via logistic regression, both confidence and causal uncertainty allowed to predict the odds ratio between probabilities of acoustical and other categorization strategies, as well as between causal and other categorizations. (Lemaitre et al. 2010)

The audibility of a sound is generally deteriorated in the presence of another sound in the spectral or temporal vicinity. Such energetic auditory masking is caused by the mechanics of the inner ear. Masking effects that cannot be explained by such peripheral limitations of the ear, i.e., where both signal and masker are well resolved by the peripheral auditory

2.1. Source identification, causal uncertainty, and informational masking

Table 2.1.: Overview of listening modes (adapted from Supper and Bijsterveld 2015).

why	how		
	<i>synthetic listening</i>	<i>analytic listening</i>	<i>interactive listening</i>
<i>monitoring listening</i>	Listening to overall features of sound for the purposes of monitoring.	Attending to specific characteristics of sound for the purposes of monitoring.	Interacting with a sound source for the purposes of monitoring.
<i>diagnostic listening</i>	Using a (quick) overall impression of a sound for the purposes of diagnosis.	Attending to specific characteristics of a sound for the purposes of diagnosis.	Interacting with a sound source for the purposes of diagnosis.
<i>exploratory listening</i>	Listening out for general impressions for the purposes of exploration.	Attending to specific features of sound for the purposes of exploration.	Interacting with the sources of a sound for the purposes of exploration.

system, are referred to as informational masking (Neuhoff 2004, pp. 199–200).

Oh and Lutfi (1999) examined the informational masking of a 1 kHz tone by brief everyday sounds. 50 sounds were selected from a sound effects CD; their almost perfect identification by listeners was proved in a pilot study. Additionally, synthesized maskers were produced by recreating the spectrum of the everyday sounds via filtered noise. In each trial, participants decided which of two stimuli contained the signal, while the signal level was varied adaptively in order to find the threshold of masking. The results revealed that everyday sounds achieve a significantly larger amount of masking (4 to 10 dB) than noise of equal power spectrum. In a follow-up experiment, however, this noise advantage was eliminated for everyday sounds that were rated as easy to recognize.

Supper and Bijsterveld (2015) argue that distinguishing between modes of listening helps to understand the active listening practices that scientists, doctors, engineers, and mechanics developed to gain knowledge on bodies, machines, and other objects of research. Their two-dimensional taxonomy of listening practices takes into account the purpose of listening as well as the ways of doing so (see Tab. 2.1). In general, practitioners shift between listening modes depending on the current task. These modes include not only passive perception but also active sonic skills such as knowledge on proper positioning of a stethoscope during auscultation.

Within the first dimension of listening, Supper and Bijsterveld (2015) distinguish between three *purposes of listening*. *Monitoring listening* refers to the monitoring of information as secondary task in the background, while performing another main

task, for example, when checking for possible malfunctions of a car while driving. *Diagnostic listening* comes into play when trying to pinpoint what exactly is wrong. It includes constant comparison of the perceived, possibly abnormal sound, with an inner reference of one or multiple stereotypical sounds, as applied, for example, in quality control. *Exploratory listening* refers to listening out for new phenomena, unprejudiced, with the objective to gain knowledge that wasn't thought of before. This includes, for example, tapping on a wall in the search for a solid spot for placing a nail.

Within the second dimension of listening, Supper and Bijsterveld (2015) distinguish between three *ways of listening*. *Synthetic listening* refers to the interpretation of auditory information as generally as possible, for example, when listening to a room full of voices or to the overall effect of a piece of music as a whole. On the contrary, *analytic listening* takes place when components of the auditory scene are identified at a finer level. This includes, for example, tuning of the focus to individual instruments of an orchestra, or to the couple whispering in the next row (Hermann et al. 2011, p. 3). While synthetic and analytic listening are usually defined as opposites, they have in common that the sound source itself is stable or unfolds at its own dynamic rules. *Interactive listening* means that the listener itself intervenes in the sounds while listening. This includes, for example, listening for changes in the sound of a car while changing gears.

In combination, the listed purposes and ways of listening lead to the 6 distinct listening modes in Tab. 2.1. While this representation suggests the selection of one specific listening mode for a given task, Supper and Bijsterveld (2015) argue that the

2. The psychophysics of auditory feedback: an annotated bibliography

full potential is unfolded by gradually shifting between those modes. The ability to shift expresses the virtuosity of sonic skills within everyday knowledge practices that go beyond the modes of listening themselves.

Summing up, in sound identification we distinguish between animate concepts and inanimate sound sources. For inanimate sounds, e.g., from a struck solid object, identification takes significantly longer (between 100 ms and 400 ms) than for animate sounds such as a dog bark (as low as 50 ms). In addition, we have a larger vocabulary for inanimate than for animate sounds. This is also expressed by the causal uncertainty which is higher with increasing typicality of the sound. Sounds that are not easy to identify (e.g., due to high causal uncertainty) mask other sounds due to informational masking by about 4 to 10 dB in addition to auditory masking. When listening to auditory augmentations, we make use of different modes of listening. In a background task, the auditory display is perceived via monitoring and interactive listening. If a certain aspect attracts the user's attention, he or she may switch to diagnostic and analytic listening to retrieve more precise information on a certain aspect of the display.

References

- Ballas, James A. (1989). *Acoustic and Perceptual-Cognitive Factors in the Identification of 41 Environmental Sounds*. AD-A214 944. Fairfax, VA: Center for Behavioral and Cognitive Studies, George Mason University.
- (1993). “Common Factors in the Identification of an Assortment of Brief Everyday Sounds.” In: *Journal of Experimental Psychology: Human Perception and Performance* 19.2, pp. 250–267.
- Giordano, Bruno L. and John McDonnell (2007). “Acoustical and conceptual information for the perception of animate and inanimate sound sources”. In: *International Conference on Auditory Display (ICAD)*. Montréal, Canada, pp. 173–180.
- Giordano, Bruno L., John McDonnell, and Stephen McAdams (June 2010). “Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds”. In: *Brain and Cognition* 73.1, pp. 7–19. DOI: 10.1016/j.bandc.2010.01.005.

Guillaume, Anne et al. (2004). “How long does it take to identify everyday sounds?” In: *International Conference on Auditory Display (ICAD)*. Sydney, Australia.

Hermann, Thomas, Andy Hunt, and John G Neuhoff (2011). *The sonification handbook*. Berlin: Logos Verlag. ISBN: 978-3-8325-2819-5.

Lemaitre, Guillaume et al. (2010). “Listener expertise and sound identification influence the categorization of environmental sounds.” In: *Journal of Experimental Psychology: Applied* 16.1, pp. 16–32. DOI: 10.1037/a0018762.

Neuhoff, John G., ed. (2004). *Ecological psychoacoustics*. Amsterdam; Boston: Elsevier Academic Press.

Oh, Eunmi L. and Robert A. Lutfi (Dec. 1999). “Informational masking by everyday sounds”. In: *The Journal of the Acoustical Society of America* 106.6, pp. 3521–3528. DOI: 10.1121/1.428205.

Supper, Alexandra and Karin Bijsterveld (June 2015). “Sounds Convincing: Modes of Listening and Sonic Skills in Knowledge Making”. In: *Interdisciplinary Science Reviews* 40.2, pp. 124–144. DOI: 10.1179/0308018815Z.000000000109.

2.2. Auditory perception of physical parameters

This section summarizes research on auditory perception of physical parameters, with a focus on the sound of rigid objects. It includes literature from earlier collections such as by Rocchesso and Fontana (2003, pp. 1–16), Carello et al. (2005), and Yost et al. (2008, pp. 13–42), as well as more recent literature until the end of year 2021.

2.2.1. Material category

Lemaitre and Heller (2012) examined the auditory discrimination between materials of cylinders under different conditions of 2 sizes and 4 manners of excitation (bouncing, hitting, rolling, scraping). The 4 materials came from 2 gross density categories: low density (plastic, wood) and high density (glass, metal). Participants judged how well each stimulus conveyed that it originated from a given *assumed* material; for all combinations of true and assumed material. They were good at discerning gross den-

sity categories (probability of superiority $P_s = 0.89$)¹ but could hardly discriminate materials within a category ($P_s = 0.46$). Discrimination was better for impacts (bounding and hitting) than for continuous excitations (hitting and rolling); however, the manner of excitation had almost no effect on the overall results. There was a strong negative correlation between reaction times and the proportion of correct answers.

Lutfi and Oh (1997) found that discrimination of material is poor due to the fact that listeners fail to make use of any information except frequency. Listeners were asked to discriminate iron from silver, steel, and copper, as well as glass from crystal, quartz, and aluminum. The synthesized sounds jittered in base frequency, amplitude, and overall decay time, to allow an estimation of the contribution of each sound parameter on the decision. For each pair of target material (iron or glass) and one of the three corresponding materials, ideal decision weights of sound parameters were computed. Overall, participants gave greatest weight to frequency, hence overstating its informative value. The effect of weighting on identification performance was highly significant and presumably responsible for the listeners' performance reduction of nearly 80% in some cases.

Giordano and McAdams (2006) let listeners identify the material of square plates struck by a steel pendulum. Plates were 2 mm thick and differed in length (between 8.66 cm and 34.64 cm) and material category (acrylic/plastic, glass, steel, walnut wood). Participants listened to the individual monophonic recordings and assigned each stimulus to one of the 4 materials. 88% of the listeners showed perfect identification of gross density categories (steel/glass vs. wood/acrylic), independently of the plate dimensions. Within these categories, materials were perceptually equivalent. The results could be predicted through multiple regression based on acoustical descriptors. The base frequency (the first peak above a certain threshold) allowed discrimination between metal and glass ($\chi^2(8) = 9.7$). For discrimination between wood and plastic, a combination of base frequency with loudness was necessary ($\chi^2(8) \leq 11.9$). In theory, base frequency and loudness together allow perfect discrimination between the 4 materials, as well as a combination of duration and spectral centroid.

¹The probability of superiority is equivalent to the area under the receiver operating characteristic (ROC) curve (AUROC or simply AUC), sometimes also called common language effect size (CL) (Ruscio 2008).

Tucker and Brown (2003) asked listeners to identify the material (metal, plastic, wood) of impacted physical plates based on audio recordings. Participants were able to perfectly identify the gross density category (metal vs. plastic and wood) but were unable to distinguish plastic and wood, independent of a plate's shape (circle, square, triangle).

Fontana et al. (2011) carried out a material identification experiment with recorded walking sounds on solid (wood, concrete) and aggregate ground (gravel, twigs). From each of the 4 natural stimuli, 6 additional variants were created by applying the temporal or spectral characteristics of the 3 other sounds, respectively. While natural stimuli were almost perfectly identified, manipulated sounds were attributed to the different materials depending on plausible combinations. In general, listeners tolerated a substitution of temporal features within, but not between categories. Spectral substitutions were generally tolerated.

McAdams et al. (2010) synthesized sounds of impacted plates on a continuous material scale between glass and aluminum by manipulating the damping interpolation parameter H (see 3.2.6 for physical explanation). Participants were found to base their judgments only on the damping-related properties, contrary to prior dissimilarity ratings where also pitch was found to be an important factor. Listeners thus tend to evaluate only the acoustic information that is reliable for the given task. The decision threshold was at $H = 0.63$; values below 0 (ultra-glass) and above 1 (ultra-aluminum) led to a convergence of the psychometric function, i.e., perfect identification of metallicity. Results were independent of the material of the striking mallet (wood/rubber).

Zhang et al. (2017) synthesized sounds of falling objects of 4 different materials (steel, ceramic, polystyrene, and wood) bouncing on a hard surface. Untrained listeners achieved accuracies around 43%.

Hermes (1998) examined auditory material identification on synthesized impact sounds varying in center frequency and time constant τ of the exponential decay. In a free identification experiment, participants labeled materials of 7×7 parameter combinations between 0.1 and 6.4 kHz and between 6.2 and 50 ms, respectively. The most frequently mentioned materials were wood ($107 \times$), metal ($74 \times$), glass ($52 \times$), plastic ($36 \times$), and rubber/skin (consolidated category combining different names). These were found to generally describe distinct regions in the 2D parameter space of center frequency and

2. The psychophysics of auditory feedback: an annotated bibliography

decay time.

These 5 material categories were used in a forced-choice experiment where participants had to select the perceived material. While identification performance for metal, wood, and glass was equally high, identification of rubber and especially plastic was inferior and differed between participants. These results are in line with the prior outcome that metal, glass and wood are well-defined perceptual categories, while the mental representation of plastic and rubber/skin is somewhat blurred.

Klatzky et al. (2000) used synthesized sounds of clamped ideal bars struck at 61% length. The stimuli were equally spread (on a logarithmic scale) over a parameter space of 5 fundamental frequencies f_1 (between 0.1 and 1 kHz) and 5 time constants τ of the decay (between 3 and 300 s). In pairwise comparisons between combinations of f_1 and τ , participants gave similarity ratings concerning the perceived material. A multidimensional scaling analysis suggested two perceptual dimensions which strongly correlated with the model parameters f_1 and τ ($r=0.96$ and 0.98 , respectively), while the spread within dimensions gave rise to a greater contribution of the decay parameter.

If starting amplitudes were jittered within 10 dB, simulating a random force of impact, a larger amplitude difference led to larger dissimilarity; this effect was stronger for decay differences than for frequency differences. In a third experiment, participants rated the similarity in object length on the same set of stimuli with variable amplitude. While the influence of decay on participants' judgments was predominant under all 3 mentioned conditions, the contribution of decay was still substantially smaller in case of length judgments.

In a fourth experiment, participants assigned each of the 25 sounds to one of 4 material categories (rubber, wood, glass, steel). On average, the probability that a stimulus was assigned to a certain material was high in the corners of the parameter space. In particular, glass was assigned to high frequency and long decay, wood to high frequency and short decay, steel to low frequency and long decay, and rubber to low frequency and short decay. The extreme combinations were even global maxima of identification probability except for steel where the corner represented the 2nd-largest local maximum. The results were predicted by the two model parameters via linear regression (on a logarithmic scale), with values of R^2 between 0.62 (glass) and 0.82 (steel).

Lutfi and Stoelinga (2010) asked participants to

discriminate between impacted bars of metal and glass (radius 3 cm, length 10 cm). Participants had to select the metal bar within pairs of stimuli. In two different conditions, sounds were synthesized according to either a viscous damping model (see Chaigne and Doutaut 1997) or a viscoelastic damping model (see Lutfi and Oh 1997); 3 partials were rendered. Analytical results and just-noticeable differences (JNDs) suggested that listeners could exploit information related to both frequency and decay in case of the viscoelastic model, but had to rely on frequency alone in case of the viscous model. These predicted decision weights were confirmed by the results of the listening test for all 3 subjects via multivariate regression.

Koumura and Furukawa (2017) examined the effect of reverberation on the identification of material via short impact sounds. They found that reverberation actually deteriorates material identification; however, after a short while, participants adapted to the reverberation and achieved similar identification rates as with the dry stimuli. It must be noted that the results varied greatly among participants. The authors suggest that there is a mechanism in the auditory system which compensates reverberation by adapting to the spectral features of the reverberant sound. In a follow-up experiment, they found, however, that there is no common mental representation of reverberation that affects all kinds of auditory perception. In contrast, learned reverberation is not transferable between gross categories of sounds. In particular, the adaptation to reverberation during speech does not help to identify a following impact sound (Koumura and Furukawa 2018).

In a material identification experiment by Traer et al. (2019), participants discriminated between metal, ceramic, wood, and cardboard based on impact sounds. With recorded sounds, participants achieved excellent performance for discrimination between gross hardness categories (metal/ceramic vs. wood/cardboard), but made errors within categories (metal vs. ceramic, wood vs. cardboard). This pattern had already been shown in prior studies and occurred similarly with synthesized sounds. Correlation between confusion matrices of real and synthetic sounds was moderately high (0.72).

Aramaki et al. (2009) used an analysis-synthesis approach to create stimuli of impact sounds from recordings of rigid objects of different material (wood, metal, glass). The resynthesized sounds were then tuned to the same chroma (the nearest octave of note *C*). In addition to the perfect material sounds, in-between sounds were synthe-

sized by morphing between model parameters. In an experiment, participants identified the material of each stimulus as fast as possible by pressing one of three buttons. An acoustical analysis of the stimuli showed that decay time alone was sufficient to discriminate between the 3 material categories. On average, participants more often categorized sounds as metal than as glass or wood. Response times were significantly slower for glass sounds than for wood and metal sounds. The in-between sounds gave a hint on what acoustical feature was linked to what material. Sounds with metal damping but glass or wood spectrum were categorized as metal. Sounds with wood damping but metal spectrum were perceived as wood, while wood damping and glass spectrum was perceived as glass. Sounds with glass damping but wood or metal spectrum led to a perception as metal. Decisions were thus made not only based on damping, but also on spectral features. This was confirmed by electrophysiological data in form of event-related brain potentials (ERPs). For sounds that were typical within their category (similar classification by more than 70% of the participants), measured ERPs at different regions suggested a correlation with temporal descriptors (decay time) and spectral descriptors (spectral centroid, etc.).

References

- Aramaki, Mitsuko et al. (2009). "Timbre Perception of Sounds from Impacted Materials: Behavioral, Electrophysiological and Acoustic Approaches". In: *Computer Music Modeling and Retrieval (CMMR)*. Vol. 5493. Springer, pp. 1–17. DOI: 10.1007/978-3-642-02518-1_1.
- Chaigne, Antoine and Vincent Doutaut (Jan. 1997). "Numerical simulations of xylophones. I. Time-domain modeling of the vibrating bars". In: *The Journal of the Acoustical Society of America* 101.1, pp. 539–557. DOI: 10.1121/1.418117.
- Fontana, Federico et al. (2011). "Auditory Recognition of Floor Surfaces by Temporal and Spectral Cues of Walking". In: *International Conference on Auditory Display (ICAD)*. Budapest, Hungary.
- Giordano, Bruno L. and Stephen McAdams (2006). "Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates". In: *Journal of the Acoustical Society of America (JASA)* 119.2, pp. 1171–1181. DOI: 10.1121/1.2149839.
- Hermes, Dik J. (1998). "Auditory material perception". In: *IPO Annual Progress Report* 33, pp. 95–102. DOI: 10.1037/e492042004-001.
- Klatzky, Roberta L., Dinesh K. Pai, and Eric P. Krotkov (Aug. 2000). "Perception of Material from Contact Sounds". In: *Presence: Teleoperators and Virtual Environments* 9.4, pp. 399–410. DOI: 10.1162/105474600566907.
- Koumura, Takuya and Shigeto Furukawa (Dec. 2017). "Context-Dependent Effect of Reverberation on Material Perception from Impact Sound". In: *Scientific Reports* 7.1. DOI: 10.1038/s41598-017-16651-4.
- (Sept. 1, 2018). "Do Speech Contexts Induce Constancy of Material Perception Based on Impact Sound Under Reverberation?" In: *Acta Acustica united with Acustica* 104.5, pp. 796–799. DOI: 10.3813/AAA.919226.
- Lemaitre, Guillaume and Laurie M. Heller (Feb. 2012). "Auditory perception of material is fragile while action is strikingly robust". In: *The Journal of the Acoustical Society of America* 131.2, pp. 1337–1348. DOI: 10.1121/1.3675946.
- Lutfi, Robert A. and Eunmi L. Oh (Dec. 1997). "Auditory discrimination of material changes in a struck-clamped bar". In: *The Journal of the Acoustical Society of America* 102.6, pp. 3647–3656. DOI: 10.1121/1.420151.
- Lutfi, Robert A. and Christophe N. J. Stoelinga (Jan. 2010). "Sensory constraints on auditory identification of the material and geometric properties of struck bars". In: *The Journal of the Acoustical Society of America* 127.1, pp. 350–360. DOI: 10.1121/1.3263606.
- McAdams, Stephen et al. (2010). "The psychomechanics of simulated sound sources: Material properties of impacted thin plates". In: *The Journal of the Acoustical Society of America* 128.3. DOI: 10.1121/1.3466867.
- Ruscio, John (2008). "A probability-based measure of effect size: Robustness to base rates and other factors." In: *Psychological Methods* 13.1, pp. 19–30. DOI: 10.1037/1082-989X.13.1.19.
- Traer, James, Maddie Cusimano, and Josh H. McDermott (2019). "A perceptually inspired generative model of rigid-body contact sounds". In: *International Conference on Digital Audio Effects (DAFx)*. Birmingham, UK.
- Tucker, Simon and Guy J. Brown (2003). "Modelling the auditory perception of size, shape and material: Applications to the classification of transient sonar sounds". In: *AES Convention*. Amsterdam, Netherlands: Audio Engineering Society.

Zhang, Zhoutong et al. (2017). "Shape and Material from Sound". In: *Conference on Neural Information Processing Systems (NIPS)*.

2.2.2. Action

Lemaitre and Heller (2012) examined the auditory discrimination between manners of excitation (actions) for cylinders of different materials and sizes. The 4 actions came from 2 gross categories: impacted (bouncing, hitting) and continuous (rolling and scraping). Participants judged how well a given stimulus conveyed that it originated from a given (assumed) action; for all combinations of true and assumed action. Participants showed almost perfect discrimination between ($P_s=0.97$) and across categories ($P_s=0.96$), independent of material and size. There was a strong negative correlation between reaction times and the proportion of correct answers.

Serafin et al. (2010) asked listeners to identify by sound if a person walks over flat ground, a bump, or a hole. Participants could successfully identify bumps and holes based on the rhythmic pattern, with between 70 % and 100 % correct identifications. Wrong answers co-occurred with low confidence ratings.

Auditory discrimination between breaking and bouncing glass objects was examined by Warren Jr. and Verbrugge (1984). For natural recordings, average correct identification reached 99 %. Sounds constructed from individual samples reached 89 % correct identification. Results did not change significantly if the first impact burst was removed from the natural and constructed sounds. Listeners are thus able to exploit the rhythmic impact pattern for discriminating between breaking and bouncing.

Zhang et al. (2017) synthesized sounds of objects falling from two different heights on a hard surface, thus creating bouncing impact patterns. Untrained listeners achieved identification accuracies around 77 %. It must be noted that the sounds always followed a pause which corresponded to the duration of falling; however, participants were not informed about this fact and thus failed to evaluate this information efficiently.

References

- Lemaitre, Guillaume and Laurie M. Heller (Feb. 2012). "Auditory perception of material is fragile while action is strikingly robust". In: *The Journal of the Acoustical Society of America* 131.2, pp. 1337–1348. DOI: 10.1121/1.3675946.
- Serafin, Stefania, Luca Turchet, and Rolf Nordahl (2010). "Do you hear a bump or a hole? An experiment on temporal aspects in the recognition of footsteps sounds". In: *International Conference on Digital Audio Effects (DAFx)*. Graz, Austria.
- Warren Jr., William H. and Robert R. Verbrugge (1984). "Auditory Perception of Breaking and Bouncing Events: A Case Study in Ecological Acoustics". In: *Journal of Experimental Psychology: Human Perception and Performance* 10.5, pp. 704–712.
- Zhang, Zhoutong et al. (2017). "Shape and Material from Sound". In: *Conference on Neural Information Processing Systems (NIPS)*.

2.2.3. Material properties

McAdams et al. (2004) used synthesized sounds of impacted bars with a tube resonator to analyze the perceptual space of physical parameters via dissimilarity ratings. Multidimensional scaling analysis showed that the dimensionality of the perceptual space was identical to that of the physical space. For bars varying only in density ρ and (frequency-independent) loss factor η , 92 % of the variance of dissimilarity ratings could be predicted by a model of two perceptual dimensions which showed strong correlation to a pair of acoustical descriptors including a damping-related one (loss factor or slope of the spectral centroid) in conjunction with base frequency f_1 . For tuned bars differing in η and variable cross section, the two perceptual dimensions explained between 85 % and 86 % of the variance, and correlated with the same set of acoustical descriptors. Dissimilarity ratings could be predicted by a combination of base frequency f_1 , (late) decay factor α_2 , and spectral centroid ($R^2 \geq 0.64$, depending on the dataset).

McAdams et al. (2010) conducted a similar experiment with synthesized sounds of thin plates that differed in damping interpolation parameter H (which blends between viscoelastic and thermoelastic damping) and longitudinal wave velocity c_L . According to dissimilarity ratings, the two perceptual dimensions were closely related to the mechanical properties ($R^2 > 0.99$ and $R^2 > 0.89$ for H and c_L ,

respectively), independent of mallet material (rubber/wood). In terms of acoustical descriptors, c_L was related to pitch while damping was associated with timbre and duration.

Lutfi and Ching-Ju Liu (2007) let participants listen to pairs of synthesized impact sounds of variable size and density under various shape and boundary conditions (bar, plate, membrane). The task was to identify the stimulus containing the smaller but more dense object (which is usually the one with higher frequency and longer decay times). Multiple regression analysis showed that listeners based their decisions mainly on one parameter (either frequency or decay) while achieving comparable sensitivity (average d' between 1.2 and 1.9), whereas an ideal ML detector incorporating both parameters would achieve sensitivities between 2.4 and 2.8.

In a follow-up study, Ching-Ju Liu and Lutfi (2009) asked percussionists, non-percussionist musicians, and non-musicians to identify the denser of two struck objects by sound. While all groups performed similarly well, percussionists slightly outperformed the others. Additional feedback between trials immediately and significantly improved the performance of all participants, effectively equalizing the differences between groups.

Roussarie et al. (1998) measured perceptual dissimilarity of synthesized sounds of impacted bars between 16 samples in the 2D parameter space of density and damping factor. By means of multidimensional scaling, the resulting 2D perceptual space was correlated with the physical parameters as well as with acoustical descriptors. Dimension 1 could be related to the decay factor ($R^2=0.97$) as well as to a linear combination of spectral centroid and the logarithm of the time constant of the decay ($R^2=0.99$). Dimension 2 could be related to material density ($R^2=0.86$) as well as to the frequency of the 2nd partial ($R^2=0.87$). Correlations between parameters were low, suggesting orthogonality.

Lutfi and Stoelinga (2010) computed hypothetical Weber fractions $\Delta\theta/\theta$ of some material constants that affect pitch, based on the JND in pitch at a hypothetical sensitivity $d'=1$. It must be noted that this does not represent an ecologically valid scenario. They examined 11 impacted bars of different material (wood, glass, aluminum, plastic) and dimensions (length between 9 and 20 cm, radius between 3 and 9 cm). For Young's modulus E , density ρ , and also viscoelastic loss factor η , their computed Weber fraction was almost identical between 0.56 % and 1.24 % ($M=0.77\%$, $SD=0.21\%$).

Traer et al. (2019) examined the discrimination

of the masses of balls dropping on a rigid object. Real sounds were recorded from small (0.7 g) and large (7.6 g) wooden balls; synthesized sounds were created according to the physical parameters of the interaction. Participants listened to pairs of stimuli, answering which of the two contained the heavier ball. They performed equally well for real and synthetic sounds, on average with around 94 % correct identifications.

References

- Liu, Ching-Ju and Robert A. Lutfi (Apr. 2009). "Identification of impact sounds by professional percussionists." In: *The Journal of the Acoustical Society of America* 125.4, pp. 2684–2684. DOI: 10.1121/1.4784258.
- Lutfi, Robert A. and Ching-Ju Liu (Aug. 2007). "Individual differences in source identification from synthesized impact sounds". In: *The Journal of the Acoustical Society of America* 122.2, pp. 1017–1028. DOI: 10.1121/1.2751269.
- Lutfi, Robert A. and Christophe N. J. Stoelinga (Jan. 2010). "Sensory constraints on auditory identification of the material and geometric properties of struck bars". In: *The Journal of the Acoustical Society of America* 127.1, pp. 350–360. DOI: 10.1121/1.3263606.
- McAdams, Stephen, Antoine Chaigne, and Vincent Roussarie (Mar. 2004). "The psychomechanics of simulated sound sources: Material properties of impacted bars". In: *The Journal of the Acoustical Society of America* 115.3, pp. 1306–1320. DOI: 10.1121/1.1645855.
- McAdams, Stephen et al. (2010). "The psychomechanics of simulated sound sources: Material properties of impacted thin plates". In: *The Journal of the Acoustical Society of America* 128.3. DOI: 10.1121/1.3466867.
- Roussarie, Vincent, Stephen McAdams, and Antoine Chaigne (1998). "Perceptual Analysis of Vibrating Bars Synthesized with a Physical Model". In: *ICA: International Congress on Acoustics*, pp. 2227–2228.
- Traer, James, Maddie Cusimano, and Josh H. McDermott (2019). "A perceptually inspired generative model of rigid-body contact sounds". In: *International Conference on Digital Audio Effects (DAFx)*. Birmingham, UK.

2.2.4. Force of impact

Liu and Lutfi (2009) compared percussionists, non-percussionist musicians, and non-musicians in their ability to identify the stimulus containing the greater force of impact in pairwise comparisons with synthesized impact sounds. Percussionists performed only slightly better than the other groups. Feedback between trials did not significantly change participants' performance.

In a more recent study, Lutfi et al. (2011) used synthesized sounds of impacted bars to examine auditory discrimination of force of impact on three groups of listeners: percussionists, non-musicians, and participants of unknown background. They were asked to identify the sound corresponding to a greater force of impact in pairwise comparisons. While individual performance varied widely between participants, percussionists generally outperformed both other groups significantly; their decision weights based on logistic multiple regression were more close to the optimal values. The worst-performing percussionists were roughly on par with the best-performing non-musicians. While providing feedback had little effect on the decision weights, it still significantly improved the performance of non-musicians, but not that of percussionists. Unknown participants' sensitivity d' was between 0.8 and 2.2 for wood (force difference $\Delta F = 4$ dB) and between 1.1 and 2.6 for iron ($\Delta F = 3$ dB).

References

- Liu, Ching-Ju and Robert A. Lutfi (Apr. 2009). "Identification of impact sounds by professional percussionists." In: *The Journal of the Acoustical Society of America* 125.4, pp. 2684–2684. DOI: 10.1121/1.4784258.
- Lutfi, Robert A., Ching-Ju Liu, and Christophe N. J. Stoelinga (Apr. 2011). "Auditory discrimination of force of impact". In: *The Journal of the Acoustical Society of America* 129.4, pp. 2104–2111. DOI: 10.1121/1.3543969.

2.2.5. Surface roughness

Lederman (1979) used physical aluminum plates of 14 cm \times 11.4 cm \times 0.5 cm that were prepared with grooves of 0.125 mm depth and width in 8 different distances (0.18 mm and between 0.25 mm and 1 mm in steps of 0.125 mm). Participants were seated with the back to the apparatus where the experimenter performed sliding movements with the

tip of the middle finger over the plate surface at constant force (28, 112, and 448 g equivalent mass). For each condition of groove distance and force, they estimated the magnitude of surface roughness by a positive nonzero number. On average, both higher force and smaller groove distance led to significantly higher perceived roughness.

References

- Lederman, Susan J (1979). "Auditory texture perception." In: *Perception* 8, pp. 93–103.

2.2.6. Pressure and speed of rubbing

In a pilot experiment, Hermes et al. (2008) found out that naive listeners were not able to distinguish between recorded physical rubbing sounds and synthesized sounds based on filtered white noise with sine-shaped envelope. In a free identification task, a majority identified the sounds as emerging from rubbing or sweeping. For different combinations of high-pass filter cutoff frequency (2, 3, and 4.5 kHz) and spectral slope ($\alpha = \{-0.8, 0.0, 0.8\}$ with amplitude proportional to f^α), participants estimated the distance of the rubbing movement (i.e., rubbing speed) as well as effort (i.e., pressure of rubbing). Overall perceived effort decreased with increasing cutoff frequency and spectral slope. If answers were freed from individual magnitude bias, then also the estimated speed significantly increased with increasing spectral slope, while cutoff frequency had no significant effect.

Hermes et al. (2009) repeated the same experiment with recorded sounds of sheets of paper gently glided over each other back and forth. The experiment used 3 fixed tempos and 3 pressures realized by a fixed number of piled up paper sheets. Perceived effort was significantly influenced by speed and pressure, as well as by their interaction. Perceived speed (freed from magnitude bias) was significantly influenced by true pressure, but not by true speed. A correlation with the acoustical descriptors of brightness and sharpness revealed that higher brightness generally increased estimated speed and decreased estimated effort, while higher sharpness increased estimated effort. Sharpness had no influence on estimated speed.

References

- Hermes, Dik J., S. Brouwer de Koning, and P. A. P. Geelen (2009). "Perception of real rubbing sounds". In: *International Haptic and Auditory Interaction Design Workshop*. Dresden, Germany, pp. 51–60.
- Hermes, Dik J. et al. (2008). "Perception of rubbing sounds". In: *International Haptic and Auditory Interaction Design Workshop*.

2.2.7. Hardness

Freed (1990) examined the perception of mallet hardness. Participants listened to recordings of metal cooking pans of 4 different sizes being struck with 6 different percussion mallets (24 combinations in total). The mallets of different materials were assigned to hardness categories: (1) metal, (2) wood, (3) rubber, (4) cloth-covered wood, (5) felt, and (6) felt-covered rubber. Participants rated perceived mallet hardness on a continuous scale between 0 (soft) and 1 (hard). Results suggested a linear relationship between hardness category and rating, irrespective of pan size, i.e., descending perceived hardness with ascending mallet number. According to a multiple regression analysis, $R^2 = 72.5\%$ of the variance could be explained through 4 acoustical descriptors: mean and slope of spectral level, mean spectral centroid, and time-weighted average (TWA) spectral centroid. Mean spectral centroid alone led to $R^2 = 56.1\%$.

Giordano et al. (2010) asked listeners to estimate the hardness of hammer and impacted plate for synthesized sounds. For hammer hardness and plate hardness individually, listeners had to identify the stimulus with the harder hammer or plate in pairwise comparisons. In general, participants appeared to give most weight to the most accurate acoustical information, despite the fact that the same discrimination performance would have been able by different weighting strategies. Participants were better in discriminating object hardness than hammer hardness.

In a second experiment, participants rated hardness on a continuous scale. While naive listeners performed equally well in both conditions (robust Spearman correlation r_s between 0.42 and 0.52), those trained participants who already absolved the first experiment achieved significantly higher performance for object hardness (r_s approx. 0.45 and 0.88, respectively); $r_s = 0$ means chance level, $r_s = 1$ means perfect performance. In the absence of feed-

back, participants seemed to focus on the most salient information. Both information accuracy and exploitability are therefore supposed to influence the perceptual criteria.

Mallet hardness was also investigated by Lutfi and Ching-Ju Liu 2007 based on synthesized sounds. The task was to identify the impact with the softer mallet within a pair of stimuli. According to multiple regression analysis, the individual listeners used different decision weights for frequency and decay to make their judgments, independent of their actual identification performance.

A follow-up study by Ching-Ju Liu and Lutfi (2009) compared percussionists, non-percussionist musicians, and non-musicians in a similar mallet hardness discrimination task. While percussionists performed only slightly better than the other groups, all participants immediately improved if feedback was provided after each trial, thus effectively equalizing the group differences.

References

- Freed, Daniel J. (Jan. 1990). "Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events". In: *The Journal of the Acoustical Society of America* 87.1, pp. 311–322. DOI: 10.1121/1.399298.
- Giordano, Bruno L., Davide Rocchesso, and Stephen McAdams (2010). "Integration of acoustical information in the perception of impacted sound sources: The role of information accuracy and exploitability." In: *Journal of Experimental Psychology: Human Perception and Performance* 36.2, pp. 462–476. DOI: 10.1037/a0018388.
- Liu, Ching-Ju and Robert A. Lutfi (Apr. 2009). "Identification of impact sounds by professional percussionists." In: *The Journal of the Acoustical Society of America* 125.4, pp. 2684–2684. DOI: 10.1121/1.4784258.
- Lutfi, Robert A. and Ching-Ju Liu (Aug. 2007). "Individual differences in source identification from synthesized impact sounds". In: *The Journal of the Acoustical Society of America* 122.2, pp. 1017–1028. DOI: 10.1121/1.2751269.

2.2.8. Impact position

Liu and Lutfi (2009) investigated the individual differences between percussionists, non-percussionist musicians, and non-musicians in discriminating between impact positions of struck membranes. Within pairs of synthesized sounds of blocks, bars,

2. The psychophysics of auditory feedback: an annotated bibliography

plates, and membranes, participants had to identify the one with the point of mallet contact closest to the center of the object. All three groups exhibited a high frequency of reverse labeling, which means that they confused excitations at the edge with those in the center. Providing feedback between trials led to immediate dramatic improvement of performance for all participants, additionally equalizing out the small differences between groups.

References

Liu, Ching-Ju and Robert A. Lutfi (Apr. 2009). "Identification of impact sounds by professional percussionists." In: *The Journal of the Acoustical Society of America* 125.4, pp. 2684–2684. DOI: 10.1121/1.4784258.

2.2.9. Point-shaped external damping

Lutfi and Liu 2007 asked listeners to identify the damping of certain modes of a membrane, evoked by light point-shaped damping in the center. Listeners had to identify the damped membrane in pairwise comparison with an undamped one. The point-shaped damping was modeled by a reduced amplitude in certain modes. Multiple regression analysis showed that all participants relied mainly on the highest frequency partial of the model, with comparable sensitivity ($d' = 1.1$ to 1.5) and only small individual differences between participants. In comparison to other studies, differences are assumed to be less likely to appear, if the relevant information is restricted only to a subset of partials.

References

Lutfi, Robert A. and Ching-Ju Liu (Aug. 2007). "Individual differences in source identification from synthesized impact sounds". In: *The Journal of the Acoustical Society of America* 122.2, pp. 1017–1028. DOI: 10.1121/1.2751269.

2.2.10. Size

Carello et al. (1998) asked participants to estimate the length of wooden rods of 1.27 cm diameter by just listening to the sound of them falling on the floor. Linear regression of perceived length to actual length showed a linear relationship with slope 0.44 and $R^2 = 0.95$ for short rods (10 cm to 40 cm), and slope 0.78 and $R^2 = 0.95$ for long rods (30 cm to 120 cm); the pooled data resulted in 0.77 slope

and $R^2 = 0.97$. According to a multiple regression analysis, the perceived length could be related to the principal rotational inertia of the rods ($R^2 = 0.97$).

Tucker and Brown (2003) asked listeners to estimate the size ratio between two plates by adjusting a visual representation of both plates on a screen. Stimuli were recordings of impacted physical plates of 3 sizes. Listeners consistently underestimated the size ratio of the plates. Gross errors were largely confusions between medium and large plates; participants performed better on metallic plates than on plastic and wood.

Lutfi and Stoeltinga (2010) computed hypothetical Weber fractions $\Delta\theta/\theta$ of length and radius of impacted bars, based on JNDs of the respective sound parameters, at a hypothetical sensitivity $d' = 1$. It must be noted that this does not represent an ecologically valid scenario. They included bars of different materials, sizes, and shapes. For length, they obtained Weber fractions between 0.0005 % and 0.0019 % ($M = 0.0010$ %, $SD = 0.0005$ %). For radius, the Weber fraction was estimated to lie between 0.28 % and 0.62 % ($M = 0.38$ %, $SD = 0.10$ %).

Grassi et al. (2013) examined the ability of listeners to identify the size of a ball by the sound it makes when dropping on a plate. Physical solid wooden balls of different diameter between 1 cm and 5 cm were dropped by the experimenter on a ceramic surface from variable height (3, 6 or 12 cm). The balls produced multiple (at least one) rebounds. Naive listeners without prior training were asked to estimate the size of each dropping ball by adjusting a 2D visual representation to scale on a computer screen. Overall, a larger true size led to a significantly larger estimated size, while absolute sizes were generally underestimated, especially smaller ones. In addition, a greater height led to significantly higher estimated size as well; however, the analysis of effect sizes revealed a stronger influence of true size than of dropping height. Participants reported that they actually recognized that there were different dropping heights. A linear regression analysis (on a logarithmic scale) for each height condition showed that estimated sizes were highly correlated with true sizes ($R^2 \geq 0.988$). It was concluded that the size information was mainly inferred from spectral centroid and intensity; the fact that both acoustical descriptors are also affected by the dropping height, already predicts a confusion. Similar results had been reported from a previous study (Grassi 2005).

The auditory perception of size and speed of

rolling balls was investigated by Houben et al. in various experiments based on recorded sounds (Houben et al. 1999; Houben et al. 2001; Houben et al. 2004; Houben et al. 2005). In general, listeners were able to correctly discriminate between ball sizes, independent of rolling speed. The estimated speed, however, was influenced by the size. This discrepancy was explained through conflicting auditory cues when discerning size and speed.

References

- Carello, Claudia, Krista L. Anderson, and Andrew J. Kunkler-Peck (May 1998). "Perception of Object Length by Sound". In: *Psychological Science* 9.3, pp. 211–214. DOI: 10.1111/1467-9280.00040.
- Grassi, Massimo (Feb. 2005). "Do we hear size or sound? Balls dropped on plates". In: *Perception & Psychophysics* 67.2, pp. 274–284. DOI: 10.3758/BF03206491.
- Grassi, Massimo, Massimiliano Pastore, and Guillaume Lemaitre (May 2013). "Looking at the world with your ears: How do we get the size of an object from its sound?" In: *Acta Psychologica* 143.1, pp. 96–104. DOI: 10.1016/j.actpsy.2013.02.005.
- Houben, Mark M. J., Dik J. Hermes, and A. G. Kohlrausch (1999). "Auditory perception of the size and velocity of rolling balls". In: *IPO Annual Progress Report* 34, pp. 86–93.
- Houben, Mark M. J., Armin Kohlrausch, and Dik J. Hermes (2001). "Auditory cues determining the perception of the size and speed of rolling balls". In: *International Conference on Auditory Display (ICAD)*. Espoo, Finland, pp. 105–110.
- (Sept. 2004). "Perception of the size and speed of rolling balls by sound". In: *Speech Communication* 43.4, pp. 331–345. DOI: 10.1016/j.specom.2004.03.004.
- (2005). "The Contribution of Spectral and Temporal Information to the Auditory Perception of the Size and Speed of Rolling Balls". In: *Acta Acustica united with Acustica* 91, pp. 1007–1015.
- Lutfi, Robert A. and Christophe N. J. Stoelinga (Jan. 2010). "Sensory constraints on auditory identification of the material and geometric properties of struck bars". In: *The Journal of the Acoustical Society of America* 127.1, pp. 350–360. DOI: 10.1121/1.3263606.
- Tucker, Simon and Guy J. Brown (2003). "Modelling the auditory perception of size, shape and material: Applications to the classification of tran-

sient sonar sounds". In: *AES Convention*. Amsterdam, Netherlands: Audio Engineering Society.

2.2.11. Shape of rigid objects

Tucker and Brown (2003) asked listeners to identify the shape (circle, square, triangle) of impacted physical plates based on sound recordings. Results showed little correspondence between actual shape and perceived shape ($\chi^2(4) = 1.33$). Perceived shape, however, interacted with material: participants often connected plastic with squares and wood with circles ($\chi^2(4) = 26.01$).

Zhang et al. (2017) synthesized sounds of falling objects of three different shape attributes ("with edge", "with curved surface", and "pointy") bouncing on a hard surface. Untrained listeners achieved identification accuracies around 70%.

Kunkler-Peck and Turvey (2000) performed experiments with physical plates of variable shape that were freely hanging in a frame, suspended with fishing line, and occluded by an acoustically transparent curtain. The plates were impacted by a steel pendulum; listeners received no prior training.

In experiment 1, participants directly estimated the absolute length and width of steel plates of identical area (0.23 m^2) and thickness (3.4 mm) but variable aspect ratio (1:1, 1.6:1, and 3.6:1) by adjusting horizontal and vertical lines in front of the curtain. In general, estimated length and width differed significantly across the true values, except between medium and long plates. Participants consistently underestimated length and width, as well as the ratios between two different true values. Estimated values could be predicted from true values via linear regression ($R^2(5) = 0.98$).

The same experiment was repeated with variable material (steel, wood, and acrylic) while all other parameters remained identical. While the overall results showed a similar pattern of proper ordering of dimensions, material significantly modulated the perceptual measures of shape. In particular, perceived dimensions of steel plates ($M = 38.8 \text{ cm}$) were significantly larger than those of wooden plates ($M = 26.4 \text{ cm}$), with acrylic in-between ($M = 34.1 \text{ cm}$). Overall root-mean-squared error was 19.3% and 16.3% of the actual length and width, respectively.

Another experiment by Kunkler-Peck and Turvey (2000) examined discrimination between plates of different shapes (circle, triangle, rectangle) but identical thickness and area. Participants were shown replica of the plates as visual reference. Overall

2. The psychophysics of auditory feedback: an annotated bibliography

identification accuracy was 58.3%; the main effect of shape was significantly above chance. In a similar experiment that varied also material (steel, wood, acrylic) in addition to shape, participants reached a comparable overall accuracy of 56.1%, in addition to almost perfect material identification.

Lakatos et al. (1997) tested listeners' ability to discriminate geometric shapes of bars with free boundary conditions in a cross-modal matching task. 12 steel bars of 30 cm length and 16 wooden bars of 75 cm length (in different combinations of discrete levels of width and thickness) were struck in the center of the largest side; the recorded sounds were used in a listening experiment. In case of steel, widths were 4, 6, and 8 cm, and thicknesses were 1, 2, 3, and 4 cm. In case of wood, widths were 4.4, 5.5, 6.8, and 9.2 cm, and thicknesses were 1.2, 1.8, 2.7, and 4.4 cm. Participants were asked to match the order of the visual representations of two bars' cross sections to the order of each pair of auditory stimuli. Materials were examined in two separate experiments. Training was provided in such that participants were able to strike 5 physical sample bars whose exact dimensions differed from those in the experiment. Overall classification performance was generally better in case of a greater ratio between width and thickness.

For metal bars, only 5 of 60 participants had overall accuracies below 75%. Percent correct answers were folded into the range between 50% (chance) and 100%, with this range being interpreted as dissimilarity score between 0 (identical) and 1 (very different). A multi-dimensional scaling analysis obtained two perceptual dimensions which theoretically allow perfect discrimination between block-like and plate-like bars with equal weighting of dimensions. These could be moderately correlated to spectral centroid and ratio between width and thickness. An acoustical analysis indicated that the width/thickness ratio is conveyed through frequency components of torsional and transverse bending modes. For wooden bars, 10 participants did not reach 75% accuracy. A slightly lower overall performance than for metal bars was attributed to the shorter decay times or order 1/10, as well as to the lack of transverse and torsional bending modes due to orthotropy.

References

Kunkler-Peck, Andrew J. and M. T. Turvey (2000). "Hearing Shape." In: *Journal of Experimental*

Psychology: Human Perception and Performance 26.1, pp. 279–294.

Lakatos, Stephen, Stephen Mcadams, and Caussé René (1997). "The representation of auditory source characteristics: Simple geometric form". In: *Perception & Psychophysics* 59.8, pp. 1180–1190.

Tucker, Simon and Guy J. Brown (2003). "Modelling the auditory perception of size, shape and material: Applications to the classification of transient sonar sounds". In: *AES Convention*. Amsterdam, Netherlands: Audio Engineering Society.

Zhang, Zhoutong et al. (2017). "Shape and Material from Sound". In: *Conference on Neural Information Processing Systems (NIPS)*.

2.2.12. Shape and volume of cavities

Rocchesso and Ottaviani (2001) asked listeners to compare impulse responses of cubic volumes to those of spherical volumes. For each pair of stimuli, participants decided if the cube was higher or lower in pitch than the sphere. Even if the task did not mention volumes at all, participants actually matched the physical volumes of the pair of cavities. In a second experiment, participants matched the pitch of an exponentially decaying sine wave to the impulse response of a sphere. Their pitch estimates followed a bimodal distribution around the two lowest partials.

In a follow-up study, Rocchesso (2001) used pitch-equalized (i.e., volume-matched) impulse responses of spheres and cubes in a shape matching experiment. As listeners are usually unfamiliar with raw impulse responses, they were used in conjunction with a snare drum pattern as sound source. Participants were able to freely train with labeled sounds of cubes and spheres whose volumes differed from those in the test. For each stimulus (2 shapes \times 5 volumes corresponding to spheres of diameter 0.3, 0.5, 0.7, 0.9, and 1 m), participants were asked to identify the shape. Overall classification was significantly above chance for both shapes, reaching up to 75% accuracy for large cubes. Accuracies were significantly higher for large than for small volumes; below a diameter of 50 cm, classification was not significantly different to chance. The best-performing participant repeatedly achieved perfect classification for large volumes, while accuracy dropped to 80% for the small ones.

The sounds were analyzed in terms of a correlogram, i.e., a representation of sound as a function of time, frequency, and periodicity, based on an auditory

model of the ear. It was found that cubes exhibit more than one vertical alignments of peaks, leading to the perception of more than one pitch. For the sphere, the peaks follow a curved pattern that is barely noticeable for small spheres, and which confirms their poor classification.

References

- Rocchesso, Davide (2001). "Acoustic Cues for 3-D Shape Information". In: *International Conference on Auditory Display (ICAD)*, pp. 175–180.
- Rocchesso, Davide and L. Ottaviani (2001). "Can one hear the volume of a shape?" In: *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*. NY, pp. 115–118. DOI: 10.1109/ASPAA.2001.969556.

2.2.13. Distance

Fontana et al. (2002) examined the perception of distance to a sound source inside a pipe of square cross-section (0.45 m width and height) and 9.5 m length. Both ends were modeled as total absorbers whereas the sides had natural absorption properties. While the source was located at the end of the pipe, the listening point varied in 10 steps from 0.42 m to the opposite end with a ratio of $\sqrt{2}$ between successive distances. The resulting binaural impulse responses were applied to a cowbell sound; the dynamic range across distances was less than 6 dB. In a magnitude estimation experiment without prior training, listeners estimated the absolute value of the first stimulus in m, and then evaluate all other stimuli relative to it. Averages were subtracted from the results to obtain a common logarithmic reference scale. Perceived vs. true distance was predicted by a linear model (on a logarithmic scale) with slope 0.61 and intercept of 0.46 m ($R^2=0.76$), inducing that sound sources close to the listener were generally overestimated. The perceived distance matched the true distance at around 1.25 m.

A similar study was carried out by Fontana and Rocchesso (2008) with a real PVC pipe of 10.2 m length and 0.3 m diameter. Participants with their head inside one end of the pipe listened to a speech sample from a moving loudspeaker within the pipe and estimated their distance to the sound source in a magnitude estimation procedure without prior training, similar to the previous study. The 6 discrete distances doubled between levels from 0.3 m to 9.75 m. The relation between perceived and true distance was best described by a power function $k\phi^a$

with $a=0.314$ and $k=2.4$ ($R^2=0.99$). The estimated distance matched the true distance at around 3.5 m. The average range of estimated distances was about 3.3 m. This compression was attributed to small intensity variations.

References

- Fontana, Federico and Davide Rocchesso (Aug. 1, 2008). "Auditory distance perception in an acoustic pipe". In: *ACM Transactions on Applied Perception* 5.3, pp. 1–15. DOI: 10.1145/1402236.1402240.
- Fontana, Federico, Davide Rocchesso, and L. Ottaviani (2002). "A structural approach to distance rendering in personal auditory displays". In: *IEEE International Conference on Multimodal Interfaces*. Pittsburgh, PA, USA: IEEE Comput. Soc, pp. 33–38. DOI: 10.1109/ICMI.2002.1166965.

2.2.14. Hollowness

Lutfi (2001) asked listeners to discriminate between hollow and solid bars based on synthesized sounds. Two different decision strategies were identified among participants: (1) evaluating frequency and decay (consistent with an analytic solution for hollowness), and (2) giving predominant weight to frequency.

References

- Lutfi, Robert A. (Aug. 2001). "Auditory detection of hollowness". In: *The Journal of the Acoustical Society of America* 110.2, pp. 1010–1019. DOI: 10.1121/1.1385903.

2.2.15. Boundary condition

Marquis-Favre and Faure (2008) synthesized sounds radiated from a rectangular plate with variable boundary conditions that were varied by a non-dimensional translational (K_t) and rotational (K_r) stiffness. The model supported the simulation of the three extreme conditions free (small K_t and K_r), simply-supported/hinged (large K_t , small K_r), clamped (large K_t and K_r), and everything in-between. In a pairwise comparison task, participants gave dissimilarity ratings. Multiple regression analysis of the dissimilarity ratings led to a 2D perceptual space; the two dimensions could be correlated to (1) the pitch of the first mode in mel ($r=-0.92$) and (2) Zwicker's loudness in sone ($r=0.97$).

References

Marquis-Favre, Catherine and Julien Faure (May 1, 2008). "Auditory Evaluation of Sounds Radiated from a Vibrating Plate with Various Viscoelastic Boundary Conditions". In: *Acta Acustica united with Acustica* 94.3, pp. 419–432. DOI: 10.3813/AAA.918050.

2.2.16. Veracity: real vs. synthetic

Lutfi et al. (2005) used recordings of impacted objects as well as a low-parameter modal synthesis model to evaluate the ability of listeners to discern between real and synthesized sounds. The model was tuned by ear via 4 physical parameters in order to generate sounds that roughly matched the recorded sounds. Participants listened to pairs of sounds via loudspeakers in a large room and were asked to identify the real recording. Neither professional musicians nor students performed significantly better than chance. A similar experiment where frequencies, decay times, and amplitudes of the model were set to random values achieved 81.5% correct classification. The same experiment with headphone presentation led to similar results. In a source identification experiment, the same levels of accuracy together with a similar pattern of errors was achieved for real and synthesized sounds. In addition, participants were unable to distinguish between different levels of complexity of the modal synthesis model. The authors conclude that a low-parameter modal model is sufficient even for running psychoacoustic experiments.

Traer et al. (2019) used a similar approach to test if listeners were able to distinguish between real recordings of rigid-body contact sounds and simulations based on source-filter models using modal synthesis. Their results showed that discrimination between real and synthetic was not significantly better from chance, even if crude simplifications were applied to the model (transient without modes, modes without transient).

References

Lutfi, Robert A. et al. (July 2005). "Classification and identification of recorded and synthesized impact sounds by practiced listeners, musicians, and nonmusicians". In: *The Journal of the Acoustical Society of America* 118.1, pp. 393–404. DOI: 10.1121/1.1931867.

Traer, James, Maddie Cusimano, and Josh H. McDermott (2019). "A perceptually inspired generative model of rigid-body contact sounds". In: *International Conference on Digital Audio Effects (DAFx)*. Birmingham, UK.

2.2.17. Summary

In summary, there are quite a lot of ambiguities when it comes to the identification of physical parameters by sound alone. Some of these are especially interesting concerning auditory augmentation. We cannot robustly discriminate between materials of similar density. For material identification, we tend to use pitch as our main source of information. We are not capable of identifying the impact position on membranes. We confuse the size and height of dropping balls, as well as size and speed of rolling balls. We are almost entirely unable to discriminate between round, square, and triangular plates of equal surface area. We cannot distinguish real recordings of impact sounds from a very simple modal synthesis.

Apart from this pessimistic view, we are actually not completely ignorant to the physical information that is encoded in sound. For size, shape, material category, and hardness of solid objects, for damping of membranes, as well as for volume and shape of cavities, our absolute judgments actually correlate to some extent with the physical reality. We can almost perfectly discriminate between gross density categories of materials (e.g., glass vs. plastic), and even within density categories (e.g., glass vs. aluminum) robust discrimination is possible via information encoded in the frequency-dependent damping law (viscoelastic vs. thermoelastic). After a short adaptation period, material identification is not even deteriorated through reverberation. Recognition of user action (e.g., hitting or scratching) and also object action (e.g., breaking or bouncing) is near perfect. The absolute length and width of rectangular plates is estimated with RMS error 19% and 16.3%.

For plausible auditory augmentation, we need to be pessimists and optimists at the same time. On the one hand, our shortcomings in the absolute identification of physical parameters result in a multitude of alternative auditory feedbacks. At the same time, we are able to (relatively) distinguish between them, which ensures a high bandwidth for encoding additional information in form of an auditory display. On the other hand, we are still able to perceive a large amount of the physical information.

This makes it possible to map digital information to physical parameters that can be perceived through sound by the user, based on experience from the everyday acoustic environment.

2.3. Perceptual resolution of sound parameters

If a physical parameter such as density or size affects the perceived sound, it must have an influence on one or more sound parameters (e.g., frequencies or decay times). Later, in Ch. 3, we will learn about this relationship in detail. Here, in this chapter, we will now review the perception of some selected sound parameters which are important for perception of physical properties of impacted rigid objects. Most interesting in this context is the just-noticeable difference (JND) which quantifies the minimum parameter difference that can be distinguished by human listeners, and thus describes the perceptual resolution.

According to a simple modal synthesis model as used in many of the psychoacoustic experiments reviewed above in Sec. 2.2 and described in detail in Ch. 3, we assume that the sound of impacted rigid objects is a sum of exponentially decaying sinusoids, corresponding to the objects' so-called modes. Each individual mode is defined by the following sound parameters: its starting amplitude, its frequency, and its decay time. The mode with lowest frequency corresponds to the object's base frequency. In a very simplified model, the modes are further filtered by a low-pass filter that expresses the hardness of object and impacting hammer, as well as by a high-pass filter which represents the frequency-dependent radiation efficiency of the object. The perception of these sound parameters is discussed below.

2.3.1. Decay times

According to Lutfi and Stoelinga (2010), the JNDs in decay time of exponentially decaying noise bursts can be applied to impacted bars. A conversion of the data by Schlauch et al. (2001) to a JND in time constant τ yields

$$\text{JND}(\Delta\tau) = 0.0177 \left(\frac{\tau}{s}\right)^{0.535} s \quad (2.2)$$

for short decay times $0.002 s \leq \tau \leq 0.02 s$. τ specifies the exponentially decaying envelope $e^{-t/\tau}$ and is assumed to be $1/5$ of the signal duration. As proposed by Lutfi and Stoelinga, a conversion of

data by Abel (1972) with the same assumption as before yields

$$\text{JND}(\Delta\tau) = 0.0356 \left(\frac{\tau}{s}\right)^{0.722} s \quad (2.3)$$

for long time constants $0.02 s \leq \tau \leq 0.2 s$. The whole valid range between 2 ms and 200 ms (T_{60} between 0.014 s and 1.4 s) thus fits 45 JNDs.

Järveläinen and Tolonen (2001) measured JNDs in decay times of synthesized plucked strings. Their reported values of Weber fractions between 25 % and 40 % support the validity of the above equations which give 35 % at 20 ms duration.

2.3.2. Amplitudes

Jesteadt et al. (1977) measured JNDs in intensity for pulsed sinusoids of 0.5 s duration. According to Lutfi and Stoelinga (2010), these can be applied to intensity discrimination of struck bars. A conversion to differential thresholds in amplitude yields

$$\text{JND}(\Delta A) = 0.680 (A/A_0)^{0.928} \quad (2.4)$$

where A_0 is the sensation level, i.e., the absolute detection threshold, given in amplitude. In other words, A/A_0 is the amplitude relative to the sensation level. With this threshold of hearing being a function of frequency (Robinson and Dadson 1956), the JND is not entirely independent of frequency. A plausible maximum range would be about 60 dB which roughly equals the amplitude range of music at 1 kHz (Fastl and Zwicker 2007, p. 17). This range can be split into 17 JNDs.

2.3.3. Missing partials

Stoelinga and Lutfi (2008) used the modal synthesis model of a rectangular, thin plate to evaluate the sensitivity to missing partials. Stimuli were constructed as a sum of exponentially decaying sinusoids, with the modal frequencies of an ideal, simply-supported plate. Decay was set so that $f \times \tau(f) = 200$, leading to a constant loss factor $\eta = 0.0016$. By varying plate surface area and aspect ratio, 9 conditions were created depending on the number of partials (11, 16, 24) and on the bandwidth (125 to 1125 Hz, 250 to 2250 Hz, 500 to 4500 Hz). The variable bandwidth can also be interpreted as variable base frequency f_1 at constant bandwidth of $f_{\max}/f_1 = 9$ or 3.17 octaves. The different numbers of partials lead to modal densities between 11 and 24 modes per kHz (low

f_1), between 5.5 and 12 kHz⁻¹ (medium f_1), and between 2.75 and 6 kHz⁻¹ (high f_1). Level was varied to suppress a possible effect of level differences. In a headphone experiment, highly practiced listeners were asked to detect the stimulus with a missing partial within pairs of stimuli (with feedback). The missing partial was randomly selected for each trial. On average, independent of base frequency, sensitivity d' was high (between 2 and 8) within the first 3, 4, or 6 partials, for 11, 16, or 24 total partials, respectively. Missing higher partials were rarely detected ($d' < 2$). An auditory model including spectral masking was able to make reasonable predictions of the sensitivity and thus of the modes that are best detected if missing.

These results cannot easily be transformed into a simple expression of JND. However, it shows that at least within the lowest 6 partials, a missing one can be easily detected by listeners.

2.3.4. Base frequency

Wier et al. (1977) measured JNDs in frequency for pulsed sinusoids of 0.5 s duration. Lutfi and Stoelinga (2010) suggested that these can be readily applied to frequency discrimination of struck bars. For moderate sensation levels of 40 dB sound pressure level (SPL), it was shown that

$$\text{JND}(\Delta f) = 10^{0.026\sqrt{f/\text{Hz}} - 0.533} \text{ Hz} \quad (2.5)$$

is valid in a frequency range between 0.2 kHz and 8 kHz which usually suffices for most rigid objects in our everyday acoustic environment. Note that this JND describes the minimum absolute frequency difference that is needed to identify the sign of the frequency difference between two sinusoids. The whole valid frequency range of the above equation fits 1452 JNDs.

Hedwig E. Gockel et al. (2007) showed that the discrimination of fundamental frequency f_1 of complex tones depends on the signal duration as well as the number of partials. In particular, the measured JNDs increased with decreasing duration; this effect, however, decreased with growing number of partials. The dominant region for pitch discrimination is thus shifted upwards to a frequency region with better frequency separation. A later study (Hedwig E Gockel et al. 2010) confirmed that the presence of partials effects discrimination of base frequency only at short durations (here 50 ms in comparison to 200 ms).

2.3.5. Frequency ratios / intervals

If the resonant frequencies are far enough apart from each other (about 10% according to Thurlow and Bernstein 1957), they are heard as individual pitches; their frequency ratios (intervals) contains information about the object's shape, e.g., the aspect ratio of a rectangular plate. Fantini and Viemeister (1987) measured the threshold in discrimination of frequency ratio of two-tone complex tones. For pairs of stimuli, participants were asked to identify the stimulus with the larger interval, whereas both complexes differed in average frequency. Frequencies were between 400 and 600 Hz, and the base frequency ratio equaled 1.25. For the two participants, the obtained JNDs were 1.65% and 0.83%, leading to an average of 1.24% of the base frequency ratio. One octave, i.e., a frequency ratio between 1 and 2, thus holds about 57 JNDs.

2.3.6. Modal density

An even stronger cue for the aspect ratio, however, might be the modal density DM which is defined as the total number of partials divided by the frequency difference between highest and lowest partial. Stoelinga and Lutfi (2011) found that the JND in modal density is about $\Delta DM = 0.3 DM$ for low to moderate levels of DM . The JND did not differ significantly neither over various bandwidths of an additionally applied band-pass filter, nor over different center frequencies of the tone complex. A plausible absolute maximum range of DM would be approximately between 0.001 (1 mode per kHz) and 0.1 (1 mode each 10 Hz). This range fits 18 JNDs.

2.3.7. Upper cutoff frequency / low-pass filtering

At high frequencies, the modal density is usually high enough to assume a dense and rather flat spectrum. In this case, we assume that the principles of perception of low-pass filtered noise can be transferred to the perception of impact sounds, at least at low damping.

Pickett et al. (1965) measured the JND in cutoff frequency f_c of a low-pass filter on broad-band random noise at different base cutoff frequencies (250, 500, 1000, 1500, and 2000 Hz). At each base cutoff frequency, the JND in f_c was measured. For normal-hearing listeners, the JNDs given as Weber fractions ($\Delta f_c / f_c$) equaled 12% at 250 Hz

and between 3% and 5% above. Unfortunately, there is no information on the type of filter that had been used.

A similar experiment was carried out by Campbell (1994) based on complex tones with 200 Hz base frequency and 25 harmonics of alternating phase ($0^\circ/90^\circ$). Stimuli had rise- and fall-times of 50 ms and 25 ms, respectively. Two base cutoff frequencies (1.2 and 2.4 kHz) and two filter orders (1 and 4, or -6 and -24 dB per octave) were examined. Independent of base cutoff frequency, JNDs were approximately 25% for 1st order and 4% for 4th-order low-pass filters. A 1st-order low-pass filter varied between 100 Hz and 10 kHz thus fits 18 JNDs. In case of a 4th-order low-pass filter, it extends to 100 JNDs.

2.3.8. Lower cutoff frequency / high-pass filtering

Low frequencies are generally poorly radiated by rigid objects, compared to high frequencies, depending on the individual physical parameters. The radiation efficiency can be roughly approximated by a high-pass filter. Contrary to the upper cutoff frequency (Sec. 2.3.7), we cannot assume that the modal density in this range suffices to transfer the results of filtered noise: depending on the object's cutoff frequency, only a few, or even only the base frequency, may be attenuated.

Some information might, however, be inferred from the perceptibility of missing partials (Sec. 2.3.3). A perfect high-pass filter basically eliminates the lowest partials. The sensitivity in detecting that one of the lowest partials is missing is high (around 8). In case of a high-pass filter of low order, the lowest partials are not missing, but just decreased in sound level. With the strong perceptual separation between the lowest partials, such high-pass filtering is thus assumed to be mainly perceived as level differences between individual partials.

2.3.9. Summary

The perceptual resolution of sound parameters which encode physical information of impacted solid objects is quite different among parameters. While differences in base frequency can be discriminated with exceptional precision (about 1500 JNDs within the whole parameter range), others such as amplitude, modal density, and low-pass filtering fit only about 18 levels of JNDs. With moderate precision,

we can discriminate decay factors (45 JNDs) and frequency ratios (57 JNDs). This information is useless until we know how exactly these sound parameters map to the physical parameters of physical objects. This will be discussed in detail in Ch. 3. The next section (Sec. 2.4) reviews literature on robotic perception of physical properties, where this relationship is exploited to automatically identify physical objects by sound.

References

- Abel, Sharon M. (Apr. 1972). "Duration Discrimination of Noise and Tone Bursts". In: *The Journal of the Acoustical Society of America* 51.4, pp. 1219–1223. DOI: 10.1121/1.1912963.
- Campbell, Shari L. (Sept. 1994). "Uni- and multi-dimensional identification of rise time, spectral slope, and cutoff frequency". In: *The Journal of the Acoustical Society of America* 96.3, pp. 1380–1387. DOI: 10.1121/1.411454.
- Fantini, DA and NF Viemeister (1987). "Discrimination of frequency ratios". In: *Auditory processing of complex sounds*, pp. 47–56.
- Fastl, H. and Eberhard Zwicker (2007). *Psychoacoustics: facts and models*. 3rd. ed. Springer series in information sciences 22. Berlin; New York: Springer. ISBN: 978-3-540-23159-2.
- Gockel, Hedwig E, Robert P Carlyon, and Christopher J Plack (2010). "Combining information across frequency regions in fundamental frequency discrimination". In: *Journal of the Acoustical Society of America (JASA)* 127.4.
- Gockel, Hedwig E. et al. (Jan. 2007). "Effect of duration on the frequency discrimination of individual partials in a complex tone and on the discrimination of fundamental frequency". In: *The Journal of the Acoustical Society of America* 121.1, pp. 373–382. DOI: 10.1121/1.2382476.
- Järveläinen, Hanna and Tero Tolonen (Nov. 1, 2001). "Perceptual Tolerances for Decay Parameters in Plucked String Synthesis." In: *Journal of the Audio Engineering Society* 49.11, pp. 1049–1059.
- Jesteadt, Walt, Craig C. Wier, and David M. Green (Jan. 1977). "Intensity discrimination as a function of frequency and sensation level". In: *The Journal of the Acoustical Society of America* 61.1, pp. 169–177. DOI: 10.1121/1.381278.
- Lutfi, Robert A. and Christophe N. J. Stoelinga (Jan. 2010). "Sensory constraints on auditory identification of the material and geometric properties of struck bars". In: *The Journal of the Acousti-*

2. The psychophysics of auditory feedback: an annotated bibliography

- cal Society of America* 127.1, pp. 350–360. DOI: 10.1121/1.3263606.
- Pickett, J. M., Robert L. Daly, and Susan L. Brand (Nov. 1965). “Discrimination of Spectral Cut-off Frequency in Residual Hearing and in Normal Hearing”. In: *The Journal of the Acoustical Society of America* 38.5, pp. 923–923. DOI: 10.1121/1.1939706.
- Robinson, D W and R S Dadson (May 1956). “A re-determination of the equal-loudness relations for pure tones”. In: *British Journal of Applied Physics* 7.5, pp. 166–181. DOI: 10.1088/0508-3443/7/5/302.
- Schlauch, Robert S., Dennis T. Ries, and Jeffrey J. DiGiovanni (June 2001). “Duration discrimination and subjective duration for ramped and damped sounds”. In: *The Journal of the Acoustical Society of America* 109.6, pp. 2880–2887. DOI: 10.1121/1.1372913.
- Stoelinga, Christophe N. J. and Robert A. Lutfi (May 2008). “Detection of missing modal frequencies”. In: *The Journal of the Acoustical Society of America* 123.5, pp. 3415–3415. DOI: 10.1121/1.2934148.
- (Nov. 2011). “Discrimination of the spectral density of multitone complexes”. In: *The Journal of the Acoustical Society of America* 130.5, pp. 2882–2890. DOI: 10.1121/1.3647302.
- Thurlow, W. R. and S. Bernstein (Apr. 1957). “Simultaneous Two-Tone Pitch Discrimination”. In: *The Journal of the Acoustical Society of America* 29.4, pp. 515–519. DOI: 10.1121/1.1908946.
- Wier, Craig C., Walt Jesteadt, and David M. Green (Jan. 1977). “Frequency discrimination as a function of frequency and sensation level”. In: *The Journal of the Acoustical Society of America* 61.1, pp. 178–184. DOI: 10.1121/1.381251.
- for such well-defined sounds as those from impacted rigid objects. Some research that targets the automated perception of material, shape, and user action is summarized below.

2.4.1. Material

Robotic perception of material by sound can be useful for numerous potential applications. Based on the groundbreaking work of Krotkov (1995), we will now briefly outline some major use cases.

Grasping. Humans integrate acoustic information when performing grasping operations, as has been shown by Sedda et al. (2011), so robots might benefit from that too.

Non-destructive evaluation and inspection. Instrument makers select wood based on acoustic inspection through percussion (e.g., Aramaki et al. 2007). In industry, it is common practice to perform quality control by acoustic measurements (e.g., Muravyev et al. 2013). This task is already an integral part of many modern manufacturing processes.

Reasoning about functionality. In addition to other senses such as vision, acoustic information permits deeper reasoning. It may, for example, reveal that a hard-heeled shoe could substitute for a hammer, even though their visual appearances differ.

Handling and recycling of waste. There exist already waste bins which detect their level by means of acoustic measurements (Zhao et al. 2017). And also automatic waste sorting benefits from acoustic analysis in addition to established visual or manual inspection (Huang and Pretz 2009; Lu et al. 2020).

Excavating and drilling. Construction workers must adapt to different materials and hidden obstacles such as piping, cabling, rocks, or roots of trees. Acoustical probing of obstacles may facilitate the selection of proper actions.

Traversing natural terrain. We constantly adapt our walking style depending on the auditory feedback from our feet (Marchal et al. 2013). Acoustic probing might help even robots to evaluate the trafficability of treacherous terrains.

Many physical properties such as temperature can be estimated directly via contact sensing. In some cases, however, a direct measurement of the physical parameter is not possible. For example, the mass and sliding friction of an object may be estimated by a “whack and watch” approach (Krotkov 1995) where the object is hit while mass and friction are derived from the visually measured movement and the corresponding kinetic equations. The “step and feel” approach measures the displacement of

2.4. Robotic auditory perception of physical parameters

Making sense of sound is not a unique feature of living beings. Even computers can do it. For speech there is not much to say: every mobile phone does it; however, the prevailing approach for retrieving information from sound recordings is to compare them to a massive dataset. These methods offer great possibilities, but ignore the fact that many information can also be extracted by applying mathematical models of the physical sound, especially

the ground in response to a constant force. For solid objects, we usually perform the “hit and listen” approach to derive information on an object’s material and inner structure that is hidden below its visible surface.

For “hit and listen”, Krotkov (1995) distinguishes between fixed-shape objects where information on spatial dimensions is already given, e.g., through visual inspection, and variable-shape objects where spatial dimensions are unknown, leading to a confound between material and spatial dimensions due to ambiguities in sound and thus causal uncertainty (Ballas and Sliwinski 1986). In a follow-up study, Krotkov et al. (1996; 1997) fitted quadratic functions to measured frequency-dependent loss factors in order to distinguish between aluminum, brass, glass, wood, and plastic, independently of the object’s variable shape (rods of 2 different lengths). Only for brass, a significant difference in the loss factor functions was found between the two lengths, basically due to only one measurable mode for the short rod.

An early attempt of recovering material properties from the sound of impacted solid objects was made by Wildes and Richards (1988). They derived how material categories such as rubber, wood, glass, and steel can be discriminated based on a measured decay factor. Krotkov (1995) experimentally validated this approach through measured decay rates of individual partials.

Zhang et al. (2017) used an analysis-by-synthesis approach based on the Metropolis–Hastings algorithm (Hastings 1970). The parameters of a physical model are estimated so that its sound matches a recording as good as possible. The model covers different sorts of objects falling on a hard surface, specified by 8 physical parameters: shape (cube, torus, tetrahedron, etc.), specific modulus E/ρ , initial height, rotation axis and rotation angle, as well as Rayleigh damping factors α_{RD} and β_{RD} . While shape and specific modulus included discrete classes (14 and 10, respectively), the other parameters were varied continuously. In case of unsupervised learning, the initial guess of the specific modulus started at approximately chance performance (10 % accuracy), but converged to 56 % after 80 Markov chain Monte Carlo (MCMC) iterations. The mean squared error (MSE) of α_{RD} decreased significantly with increasing iterations, while no improvement was found for the MSE of β_{RD} . With higher amount of supervision, classification accuracies generally improved up to near perfect performance with fully supervised learning. In the same classification task, untrained

humans generally achieved accuracies similar to the unsupervised algorithm after 10 to 30 iterations (see also Sec. 2.2). Both humans and algorithm occasionally confused steel with ceramic and ceramic with polystyrene due to the similar damping and specific modulus.

Tucker and Brown (2003) implemented a real-time capable model for material classification based on decay rates by using an auditory model. Within overlapping signal frames, Hilbert envelopes were taken for each band of a gammatone filterbank. For channels containing significant acoustic components, decay rates were estimated. Per-frame median decay rates were averaged (weighted by frame energy) to form a scalar value for material classification. The model achieved a good match with the experimental data by human listeners ($R^2 = 0.69$, see also Sec. 2.2.1) when distinguishing between materials (metal, plastic, and wood) of plates of different sizes.² In an application on transient sonar sounds, the same model achieved a true positive rate of 70 % with no false positives.

References

- Aramaki, Mitsuko et al. (Apr. 2007). “Sound quality assessment of wood for xylophone bars”. In: *The Journal of the Acoustical Society of America* 121.4, pp. 2407–2420. DOI: 10.1121/1.2697154.
- Ballas, J. A. and M. J. Sliwinski (Nov. 1, 1986). *Causal Uncertainty in the Identification of Environmental Sounds*. ONR-86-1. Fort Belvoir, VA: Defense Technical Information Center. DOI: 10.21236/ADA175228.
- Hastings, W. K. (Apr. 1, 1970). “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1, pp. 97–109. DOI: 10.1093/biomet/57.1.97.
- Huang, Jiu and Thomas Pretz (2009). “Feasibility of Acoustic Sorting for Black Materials in Solid Waste Processing”. In: *Waste-to-Resources International Symposium*, pp. 433–453.
- Krotkov, Eric (1995). “Robotic perception of material.” In: *IJCAI*, pp. 88–95.
- Krotkov, Eric, Roberta Klatzky, and Nina Zumel (1996). “Analysis and synthesis of the sounds of impact based on shape-invariant properties of materials.” In: *International Conference on Pattern Recognition (ICPR)*. Vol. 1. IEEE, pp. 115–119.

²The results of the same listening experiment in an underwater condition could be predicted with similar coefficient of determination ($R^2 = 0.62$).

2. The psychophysics of auditory feedback: an annotated bibliography

Krotkov, Eric, Roberta Klatzky, and Nina Zumel (1997). "Robotic perception of material: Experiments with shape-invariant acoustic measures of material type". In: *Experimental Robotics IV*. Vol. 223. London: Springer-Verlag, pp. 204–211. DOI: 10.1007/BFb0035211.

Lu, Gang et al. (Oct. 2020). "One-dimensional convolutional neural networks for acoustic waste sorting". In: *Journal of Cleaner Production* 271. DOI: 10.1016/j.jclepro.2020.122393.

Marchal, Maud et al. (2013). "Multimodal Rendering of Walking Over Virtual Grounds". In: *Human Walking in Virtual Environments*. Ed. by Frank Steinicke et al. New York, NY: Springer New York, pp. 263–295. DOI: 10.1007/978-1-4419-8432-6_12.

Muravyev, V. V., O. V. Muravyeva, and E. N. Kokorina (Jan. 2013). "Quality control of heat treatment of 60C2A steel bars using the electromagnetic-acoustic method". In: *Russian Journal of Nondestructive Testing* 49.1, pp. 15–25. DOI: 10.1134/S1061830913010075.

Sedda, Anna et al. (Mar. 2011). "Integration of visual and auditory information for hand actions: preliminary evidence for the contribution of natural sounds to grasping". In: *Experimental Brain Research* 209.3, pp. 365–374. DOI: 10.1007/s00221-011-2559-5.

Tucker, Simon and Guy J. Brown (2003). "Modelling the auditory perception of size, shape and material: Applications to the classification of transient sonar sounds". In: *AES Convention*. Amsterdam, Netherlands: Audio Engineering Society.

Wildes, Richard P. and Whitman A. Richards (1988). "Recovering material properties from sound." In: *Natural computation*, pp. 356–363.

Zhang, Zhoutong et al. (2017). "Shape and Material from Sound". In: *Conference on Neural Information Processing Systems (NIPS)*.

Zhao, Yiran et al. (Sept. 11, 2017). "VibeBin: A Vibration-Based Waste Bin Level Detection System". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3, pp. 1–22. DOI: 10.1145/3132027.

2.4.2. Shape

Zhang et al. (2017) synthesized sounds of falling objects of three different shape attributes ("with edge", "with curved surface", and "pointy") bouncing on a hard surface. An unsupervised analysis-by-synthesis approach based on the Metropolitan–Hastings algorithm was applied (Hastings 1970). Starting at

approximately chance performance (8%), after 80 iterations, it achieved between 54% (unsupervised) and 62% (weakly supervised). With fully supervised learning, accuracy was almost perfect.

References

Hastings, W. K. (Apr. 1, 1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109. DOI: 10.1093/biomet/57.1.97.

Zhang, Zhoutong et al. (2017). "Shape and Material from Sound". In: *Conference on Neural Information Processing Systems (NIPS)*.

2.4.3. Action

Alonso-Martín et al. (2017) discriminated between 4 types of human touches (stroke, tap, slap, tickle) on a rigid surface based on acoustic sensing through a contact microphone. They recorded a total of 1981 human touches from 25 different users, which was split into a training set (70%) and a testing set 30%). A total of 126 different machine learning classifiers working on pre-computed acoustical descriptors were compared on the same dataset. The best performance was obtained by a logistic model trees (LMT) classifier which achieved 82% correct classifications. This accuracy seems incredibly low compared to the near perfect classification by humans (see also Sec. 2.2.2); however, no listening test has been performed with the exact same types of actions.

References

Alonso-Martín, Fernando et al. (May 16, 2017). "Detecting and Classifying Human Touches in a Social Robot Through Acoustic Sensing and Machine Learning". In: *Sensors* 17.5. DOI: 10.3390/s17051138.

2.4.4. Summary

When trying to make use of auditory augmentations, we are mainly using the "hit and listen" approach of auditory knowledge-making—at least for strict-sense auditory augmentations of solid objects. Algorithms for robotic perception help us to better understand how physical parameters are estimated on the basis of perceived sound parameters. There are two main types of such algorithms, both utilizing

acoustical descriptors of a recorded sound. Model-based robotic perception algorithms derive physical parameters from sound parameters through a model that seeks to describe this relationship. It can be regarded as the inverse of physical modeling for sound synthesis. Data-based robotic perception algorithms compare the extracted features of a sound with a large dataset by applying machine learning techniques. Knowledge of the underlying mechanism is not necessary in this case. It seems that most research focuses on data-based methods due to their general applicability and superior results. The discussed model-based algorithms, however, are not making use of their full potential. They mainly incorporate quite rough statistical measures of sound, without making use of the sound parameters discussed in Sec. 2.3. We argue that more detailed models of the underlying physical processes may lead to significant improvements. This hypothesis is tested in Sec. 5.1.

2.5. Multisensory feedback and the integration of auditory, visual, and haptic information

We usually merge the information from various sensory modalities such as sound, vision, and haptics. In the scope of this document we refer to haptics as an umbrella term that actually includes all sorts of haptic, kinesthetic, cutaneous, tactile, and force feedback sensation. According to McGee et al. (2001), the information from different senses can be integrated in three different ways, each with the potential to affect perception and the resulting interaction.

In case of conflicting information from multiple modalities, the resulting multisensory perception may be distorted, completely lost, or changed in some rather unpredictable way.

In case of redundant, i.e., identical information from multiple modalities, only one modality might actually be processed, depending on personal ability, preference, or the nature of the task. Alternatively, the mental representation of the information may be increased by the congruent information, which may lead to increased confidence or reliability.

If the information from different senses is complementary, the combined information might be more than just the sum of the individual parts. It may

increase the quality or quantity of information in a way that is not possible with either of the unisensory percepts alone.

In the following, we review literature on multisensory perception of physical parameters, with a focus on auditory feedback from interaction with rigid objects. Only sensor combinations including sound are covered, i.e., auditory-haptic, auditory-visual, and auditory-visual-haptic perception.

2.5.1. Latency and temporal resolution

Yang and Yeh (2014) showed that auditory cues facilitate a visual detection only in case of spatial congruency between the auditory and visual stimulus. Participants were asked to react on a visual cue by pressing a button under 3 conditions: with no sound, with spatially congruent sound (loudspeaker), and with spatially incongruent sound (headphones). The addition of auditory cues to visual presentation led to significantly lower reaction times only in case of spatially congruent presentation.

Götzen and Rocchesso (2007) examined the speed/accuracy trade-off in an auditory tuning task with the metaphor of a glissando performed by a violin player following a certain tempo. Participants had to reproduce a certain glissando by controlling a variable sine tone via their hand position under different conditions of frequency interval and glissando rate. Prior to each trial, the start and end frequency was displayed as tones of constant frequency, as well as the whole glissando. Collected movement data was analyzed with respect to theory of human target acquisition based on Fitt's law and Schmidt's law. In the ISO standard version of Fitt's law, the movement time MT is derived from the distance D of a movement between start and target and from the target width W via

$$MT = a + b \log \left(\frac{D}{W} + 1 \right) \quad (2.6)$$

with regression coefficients a and b . The other way around, Schmidt's law predicts the standard deviation in endpoint coordinates W_e as a function of velocity, i.e., distance D over time MT :

$$W_e = a + b \frac{D}{MT} \quad (2.7)$$

Results indicated two different behaviors. At high speed, participants tended to stop the glissando before reaching the target due to the temporal constraint. At low speed, they tended to stop after

2. *The psychophysics of auditory feedback: an annotated bibliography*

reaching the target. In addition, measured standard deviations actually decreased with increasing speed. The absolute frequency range had no significant influence on the results. While Schmidt's law was derived mainly based on visual feedback, the results indicate that it doesn't hold for auditory feedback where the exact opposite was found: higher speed actually led to higher accuracy.

Perceptual synchrony between auditory and haptic feedback was examined by Adelstein et al. (2003) with the true haptic feedback of a reflex hammer, but synthesized auditory feedback at variable latency. Auditory feedback mimicked a struck wooden idiophone with 1 ms attack and variable decay between 1 and 200 ms, and was presented via headphones together with pink noise masking the original auditory feedback. In an adaptive procedure, participants were asked to identify the stimulus with lower latency between auditory and haptic feedback, while one stimulus had only the system's baseline latency, and the other included variable additional latency. Based on the individual participants' results, psychometric functions were fitted, allowing the computation of the point of subjective simultaneity (PSS) and the JND. Pooled over participants, the JND was 24.2 ms, while the PSS was not significantly different from zero. The decay time of auditory feedback had no significant influence on the results. It must be noted that the given system latency of 7 ms was already larger than the latency of the true physical auditory feedback (sound travels 1 m in only 3 ms) and thus allowed no negative latency which would have been necessary for a measurement of the PSS. Individual participants' JNDs reached down to 5 ms and 75% thresholds of 8 ms. The authors infer that auditory-haptic latency perception might be more sensitive than unisensory tactile temporal separation (10 ms to 30 ms).

Which kind of sensory feedback is more important for musicians trying to play along a metronome? Auditory or the tactile? This question was investigated by Dahl and Bresin (2001). Musically trained participants used an electronic percussion instrument (Max Mathews' Radio Baton) with the task to synchronize their strokes with a metronome running at 120 bpm, while the latency of the auditory feedback was gradually increased every 15 ms. A linear regression on the time difference between metronome click and impact showed that participants effectively counteracted the latency by a negative delay on their impacts. In a follow-up experiment, a more precise drum pad was used as interface. The standard deviation of inter-onset intervals increased with increasing

latency. The authors suggest a possible threshold of latency from where on participants had difficulties in keeping a steady rhythm (between 40 ms and 55 ms). Below that threshold, they were not even aware of the latency and their own adjustment to it. The threshold was interpreted as the break point where participants switched strategy from closed-loop adjustment to their internal representation of tempo, ignoring the auditory feedback.

DiFilippo and Pai (2000) explored perceptual thresholds of auditory-haptic latency with the goal to testify the potential of a novel auditory-haptic interface. A pantograph haptic device was used in conjunction with modal synthesis to render continuous contact interactions. The interface simulated a virtual wall, while participants were asked to decide whether auditory feedback preceded the haptic feedback or vice versa. The latency between auditory and haptic feedback was ± 2 ms for one stimulus; the other had no relative latency. The baseline system latency equaled 0.5 ms. There were no distinct visual cues; however, participants were not blindfolded. The measured discrimination results were not significantly different from chance. Sounds with long decay were often perceived as preceding the haptic stimulus, while sounds with short decay were perceived as lagging the haptic stimulus.

Fujisaki and Nishida (2009) compared the temporal resolution of synchrony perception across the three different bimodal combinations of auditory, haptic, and visual feedback. They found that this resolution was similar for auditory-visual and visual-tactile (4 Hz for repetitive-pulse stimuli), but significantly higher in case of auditory-tactile feedback (10 Hz). They used repetitive-pulse trains constructed from visual blobs on screen, auditory white noise bursts, and tactile pulses at the index finger as stimuli. For each stimulus, participants decided if there was synchrony or asynchrony between sensory modalities. In a follow-up experiment, participants were judged if the stimuli from two sensory modalities were simultaneous or not. The results were similar to those in the synchrony/asynchrony experiment, suggesting a superiority of auditory-tactile temporal resolution with respect to auditory-visual and visual-tactile perception.

Harrar and Harris (2008) measured the point of subjective simultaneity (PSS) as well as JNDs in relative latency for the three bimodal combinations of audition, vision, and touch. After exposure to an auditory-visual stimulus with sound lagging by 100 ms, the auditory-visual PSS shifted by 32 ms resulting in a necessary visual precedence of -9 ms.

2.5. Multisensory feedback and the integration of auditory, visual, and haptic information

The same effect was measured after exposure to a visual-tactile stimulus with touch leading by -100 ms, inducing a shifted PSS by 30 ms, again resulting in a necessary precedence of light. After exposure to an auditory-tactile stimulus with sound lagging by 100 ms, the JND of latency between audition and vision significantly decreased from 76 ms to 56 ms. Other combinations had no significant effect on PSS and JND. The authors argue for an adaptable mechanism of simultaneity constancy for auditory-visual feedback. They suggest that “the neural correlates of bimodal auditory-visual processing are re-synchronized by a separate, more flexible simultaneity constancy mechanism” than the visual-tactile or auditory-tactile processing systems.

Jack et al. (2016) investigated the effect of auditory latency on the perceived quality of musical instruments, as well as on performer interaction (see also Jack et al. 2018). Musically-trained participants improvised on an electronic percussion instrument with variable latency, based on the Bela platform (McPherson and Zappi 2015). In a first experiment, participants made relative quality estimations in pairwise comparisons, where one condition was the reference with zero latency, the other had an additional latency of either 10 ms, 20 ms, or 10 ms with ± 3 ms jitter. Average quality judgments were significantly higher for 10 ms (low latency) than for 20 ms (high latency). Perceptual quality at jittered low latency was not significantly different to high latency, suggesting that jitter has as much a negative influence on perceived quality as constant latency. In a second experiment, participants were asked to synchronize to a metronome beat as well as to freely improvise to a backing track. For each strike, the synchronization error, i.e., the strike onset relative to the metronome onset, was measured. Overall, mean synchronization error did not change significantly with variable latency. Furthermore, contrary to the observations by Dahl and Bresin (2001), participants continued to sync with tactile feedback rather than adjusting to the latency of sound. This may be attributed to the relatively small range of latency used in this study. The authors argue that digital musical instruments should aim for a latency of 10 ms or less with ± 1 ms jitter, in accordance with Wessel and Wright (2002).

Levitin et al. (2000) evaluated auditory-visual and auditory-tactile simultaneity by using an electronic percussion instrument based on Max Mathew’s Radio Baton. The two different bimodal conditions were tested simultaneously, with one blindfolded participant actively striking (auditory-tactile), and

another participant passively observing through a glass window. In both conditions, sound was presented via headphones with variable latency. For each trial, participants were asked if the two sensory modalities were synchronous or asynchronous, together with a confidence rating. On average, actors detected auditory-tactile asynchrony at -25 and 42 ms (75 % threshold), while observers detected auditory-visual asynchrony at -41 and 45 ms, with a point of subjective synchrony roughly around zero. The authors infer that the small simultaneity thresholds compared to previous studies are attributed to the greater ecological validity.

Mäki-Patola and Hämäläinen (2004) examined the effect of auditory latency on the playing of continuous sound instruments, represented by a conventional and a virtual reality theremin. In both conditions, participants performed a pitch matching task followed by the challenge to play along a given melody. Both tasks included prior training without additional latency, while the actual test was done at different levels of additional latency. On average, the time that participants needed to match a given pitch increased by factor 5 of the introduced latency, suggesting an accumulating feedback latency over the whole task. When playing along the sample melody, a latency up to 120 ms had no significant effect on timing errors (average increase of less than 20 %), while at larger latencies, errors increased up to 80 % at a latency of 240 ms.

Altinsoy (2012) provide extensive guidelines for high-quality auditory-tactile virtual environments. They suggest auditory-tactile asynchrony to be below 10 ms, and a point of subjective simultaneity (PSS) at an audio delay of 8 ms.

Summing up, the literature agrees on the fact that the latency between an action and its resulting sound is imperceptible at 10 ms or less with a maximum tolerable jitter of 1 ms. It must be noted that we are accustomed to perceive a certain delay between haptic and auditory feedback. This point of subjective simultaneity between auditory and haptic stimuli is about 8 ms. Due to sound playback via headphones, we would expect a value around 2 ms to compensate the skipped latency of sound propagation; however, the measured PSS is largely exceeding it. In order to accomplish the low round-trip latency that is necessary for plausible auditory augmentation systems, specialized tools are necessary: either strong computers with expensive professional audio interfaces or embedded systems such as the Bela platform (McPherson et al. 2016).

References

- Adelstein, Bernard D. et al. (2003). "Sensitivity to Haptic-Audio Asynchrony". In: *International Conference on Multimodal Interfaces*. Vancouver, Canada: ACM, pp. 73–76.
- Altinsoy, M Ercan (2012). "The Quality of Auditory-Tactile Virtual Environments". In: *Journal of the Audio Engineering Society* 60.1, pp. 38–64.
- Dahl, Sofia and Roberto Bresin (2001). "Is the player more influenced by the auditory than the tactile feedback from the instrument." In: *International Conference on Digital Audio Effects (DAFx)*. Limerick, Ireland, pp. 194–197.
- DiFilippo, Derek and Dinesh K Pai (2000). "The AHL: An Audio And Haptic Interface For Contact Interactions". In: *ACM symposium on User interface software and technology*, pp. 149–158.
- Fujisaki, Waka and Shin'ya Nishida (Sept. 2009). "Audio-tactile superiority over visuo-tactile and audio-visual combinations in the temporal resolution of synchrony perception". In: *Experimental Brain Research* 198.2, pp. 245–259. DOI: 10.1007/s00221-009-1870-x.
- Götzen, Amalia de and Davide Rocchesso (2007). "The speed accuracy trade-off through tuning tasks." In: *International Conference on Enactive Interfaces*. Grenoble, France, pp. 81–84.
- Harrar, Vanessa and Laurence R. Harris (Apr. 2008). "The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity". In: *Experimental Brain Research* 186.4, pp. 517–524. DOI: 10.1007/s00221-007-1253-0.
- Jack, Robert H., Tony Stockman, and Andrew P. McPherson (2016). "Effect of latency on performer interaction and subjective quality assessment of a digital musical instrument". In: *Audio Mostly*. Norrköping, Sweden: ACM Press, pp. 116–123. DOI: 10.1145/2986416.2986428.
- Jack, Robert H. et al. (Sept. 2018). "Action-sound Latency and the Perceived Quality of Digital Musical Instruments: Comparing Professional Percussionists and Amateur Musicians". In: *Music Perception* 36.1, pp. 109–128. DOI: 10.1525/mp.2018.36.1.109.
- Levitin, Daniel J et al. (2000). "The Perception of Cross-Modal Simultaneity". In: *AIP Conference Proceedings*. Vol. 517. American Institute of Physics, pp. 323–329.
- Mäki-Patola, Teemu and Perttu Hämäläinen (2004). "Effect of Latency on Playing Accuracy of Two Gesture Controlled Continuous Sound Instruments Without Tactile Feedback". In: *Conference on Digital Audio Effects (DAFx)*. Naples, Italy, pp. 11–16.
- McPherson, Andrew P., Robert H. Jack, and Giulio Moro (2016). "Action-Sound Latency: Are Our Tools Fast Enough?" In: *New Interfaces for Musical Expression (NIME)*. Brisbane, Australia.
- McPherson, Andrew P. and Victor Zappi (2015). "An Environment for Submillisecond-Latency Audio and Sensor Processing on BeagleBone Black". In: *AES Convention*. Warsaw, Poland: Audio Engineering Society.
- Wessel, David and Matthew Wright (Sept. 2002). "Problems and Prospects for Intimate Musical Control of Computers". In: *Computer Music Journal* 26.3, pp. 11–22. DOI: 10.1162/014892602320582945.
- Yang, Yung-Hao and Su-Ling Yeh (Apr. 2014). "Unmasking the dichoptic mask by sound: spatial congruency matters". In: *Experimental Brain Research* 232.4, pp. 1109–1116. DOI: 10.1007/s00221-014-3820-5.

2.5.2. Auditory-haptic feedback

According to Papetti (2013), auditory and haptic feedback is especially important in human-computer interaction. The feedback path informing about the machine's state, however, is still defective in many nowadays human-machine interfaces. The lack of proper auditory-haptic feedback often leads to a broken interaction loop. Papetti argues that auditory and haptic feedback are tightly coupled due to their common physical nature: mechanical vibrations. Applications might even already benefit from a haptic feedback that is driven by the exact same signal as the auditory feedback. In general, the combination of auditory and haptic feedback leads to an increased response as compared to unisensory auditory or haptic feedback alone. This section summarizes the effects of sensory integration and interaction between both sensory modalities, with a focus on the perception of physical properties.

2.5.2.1. Influence on walking style

Bresin et al. (2010) equipped participants with shoes that captured foot pressure and played back synthesized footstep sounds based on a real-time sound synthesis model. If the sound exhibited higher spectral centroid (e.g., when simulating iced snow), walkers used a more active walking style, i.e., faster pace, than when playing sounds with low spectral cen-

troid (e.g., simulating muddy ground). Moreover, harder texture sounds led to a more aggressive walking style than softer sounds. They concluded that there exist sounds that bring people to walk faster, independently of their emotional intention.

Papetti and Fontana (2012) used a similar system with the capability of additional auditory as well as tactile cues. They showed that illusory tactile cues can be induced by low-frequency auditory cues. In addition, they were able to manipulate participants' gait cycle by means of auditory augmentation.

Turchet and Bresin (2015) asked participants to walk in 5 different walking styles (sad, happy, tender, aggressive, neutral) on a sonically augmented floor which simulated 4 different ground materials (metal, wood, gravel, snow). In general, participants used different speed and impact force for each emotion, independent from the simulated surface material. The neutral walking style was characterized by average values of speed and impact force. Participants had different strategies to cope with the sound, mainly due to a different degree of involvement in the simulation. Some participants tried to ignore the sound to prevent themselves from being manipulated, while others overacted to it.

Maculewicz et al. (2015) showed that not only footstep sounds but also soundscapes have an effect on the preferred walking pace on an aerobic stepper.

References

- Bresin, Roberto et al. (2010). "Expressive sonification of footstep sounds." In: *Interactive Sonification Workshop (ISon)*. Vol. 2010. Stockholm, Sweden, pp. 51–54.
- Maculewicz, Justyna, Cumhuri Erkut, and Stefania Serafin (2015). "An investigation on the influence of soundscapes and footstep sounds in affecting preferred walking pace." In: *International Conference on Auditory Display (ICAD)*. Graz, Austria.
- Papetti, Stefano and Federico Fontana (2012). "Effects of Audio-Tactile Floor Augmentation on Perception and Action During Walking: Preliminary Results". In: *Sound and Music Computing Conference (SMC)*. Copenhagen, Denmark, pp. 17–22.
- Turchet, Luca and Roberto Bresin (Apr. 1, 2015). "Effects of Interactive Sonification on Emotionally Expressive Walking Styles". In: *IEEE Transactions on Affective Computing* 6.2, pp. 152–164. DOI: 10.1109/TAFFC.2015.2416724.

2.5.2.2. Material category

Turchet et al. (2010) used simulated auditory and haptic feedback, realized with augmented shoes, to examine the effect of congruent and incongruent auditory-haptic cues on the perceived ground material while walking. Actual materials consisted of wood, metal, gravel, and snow, while the participants' free identifications were gathered in groups of solid, aggregate, liquid, and unknown material. Independent of the haptic stimulus, wood and metal were mainly identified as solid material (almost never as composite), gravel and snow were never identified as solid, but almost uniquely as aggregate. The authors infer a perceptual dominance of audition over haptics for ground material identification while walking. Percentages of perceptual dominance were derived to lie between 60 % (auditory metal, haptic snow) and 100 % (auditory gravel, haptic wood).

Nordahl et al. (2011) examined how the context (in form of a soundscape) modulates the perceived material of the floor. In a first experiment, 13 different surfaces (e.g., sand, gravel, snow, leaves, wood, metal) were simulated by real-time sound synthesis controlled via contact microphones attached to a medium-density fiberboard (MDF) plate. Participants walked on the plate and were asked to identify the ground material from a list containing the 13 stimuli as well as "don't know". Recognition rates were generally high for materials such as snow, creaking wood, gravel, and metal, and low for surfaces such as pebbles, grass, and non-creaking wood. High perceptual similarity appeared together with strong physical similarities, e.g., between wood and concrete or dry leaves and forest underbrush. In a second experiment by Nordahl et al., ground materials were presented in combination with different soundscapes. Especially for those ground materials that were barely recognized, the addition of a congruent soundscape led to higher recognition rates. In case of incongruent soundscapes, responses divided between correct identification of the material and a plausible compromise in accordance with the soundscape (e.g., underbrush at a ski slope interpreted as frozen snow).

Pra et al. (2020) recorded the auditory and haptic feedback of a ping-pong ball bouncing on a plate made of wood, plastic, and metal. Haptic feedback was recorded with an accelerometer and played back with a structure-borne exciter through a suspended glass plate that participants of the experiment had to touch from underneath. Auditory feedback was recorded with a microphone and

2. The psychophysics of auditory feedback: an annotated bibliography

displayed via headphones. Under different sensory conditions (unisensory auditory, unisensory haptic, and bimodal auditory-haptic), blindfolded participants had to identify the material of the plate (wood, plastic, or metal). Participants received prior training with the original full-modal materials and physical interaction with the pin-pong ball. In general, unisensory auditory feedback achieved significantly higher recognition rates than unisensory haptic feedback, independent of material. Similarly, bimodal feedback led to significantly better identification than auditory feedback for wood and plastic, but not for metal. The bimodal presentation, however, didn't achieve the almost perfect identification rates of a control group that performed the same experiment with the original physical materials. The results suggest that auditory feedback dominates the perception of material in case of stimuli caused by an impact. The same experiment was repeated with incongruent combinations of auditory and haptic feedback. In this case, identifications were still dominated by auditory feedback, but biased towards the material of the haptic stimulus. Pra et al. argue that tactile feedback becomes more important in case of incongruent sensory information.

References

- Nordahl, Rolf, Luca Turchet, and Stefania Serafin (Sept. 2011). "Sound Synthesis and Evaluation of Interactive Footsteps and Environmental Sounds Rendering for Virtual Reality Applications". In: *IEEE Transactions on Visualization and Computer Graphics* 17.9, pp. 1234–1244. DOI: 10.1109/TVCG.2011.30.
- Pra, Yuri De et al. (May 29, 2020). "Does It Ping or Pong? Auditory and Tactile Classification of Materials by Bouncing Events". In: *ACM Transactions on Applied Perception* 17.2, pp. 1–17. DOI: 10.1145/3393898.
- Turchet, Luca et al. (2010). "Conflicting Audio-haptic Feedback in Physically Based Simulation of Walking Sounds". In: *Lecture Notes in Computer Science*. Vol. 6309. Springer Berlin Heidelberg, pp. 97–106. DOI: 10.1007/978-3-642-15841-4_11.

2.5.2.3. Material properties

Kjøer et al. (2007) and Serafin et al. (2007) investigated the influence of auditory and haptic cues on the perception of surface roughness by using a Phantom Omni haptic device in conjunction with

real-time sound synthesis for rubbing and tapping sounds. Results showed that auditory feedback improved accuracy in roughness estimation. Furthermore, auditory cues were able to alter perceived roughness in case of constant haptic roughness. The effect of auditory cues diminished if participants received prior training with haptic cues only.

McGee et al. (2002) examined the effect of conflicting, redundant, and complementary information from auditory and haptic cues on the perception of surface roughness (see also McGee et al. 2001). Haptic cues were displayed by means of a Phantom force-feedback device simulating a sine-shaped rigid wall; auditory cues were MIDI³ notes triggered near the peak of every bump. The experiment was performed under 3 conditions: unisensory haptic, multisensory congruent, and multisensory incongruent where the underlying spatial texture for auditory feedback was 20% finer (more auditory bumps) than that for the haptic feedback. In each trial, participants decided which of the two stimuli was the rougher, with the option of equality between both. On average, a higher haptic frequency led to significantly higher perceived roughness, with no significant difference between conditions. The likelihood that two haptically identical textures were perceived as identical was significantly lower for the multimodal conditions than for unisensory haptic feedback. This effect was even stronger for incongruent in comparison to congruent multimodal feedback. Furthermore, the likelihood that identical haptic frequencies were judged as different was significantly higher for multimodal than for unisensory feedback. McGee et al. suggest that the addition of auditory feedback is able to modulate perceived haptic roughness.

Furfaro et al. (2013) detected finger tapping on a wooden table through a contact microphone in order to play different types of pre-recorded tapping sounds via headphones (see also Furfaro et al. 2015). Recorded sounds included tapping on a cardboard box in three different levels of strength (weak, medium, strong). The sound was presented to blindfolded participants via headphones in combination with pink background noise in order to mask the original auditory feedback. An additional virtual condition without haptic feedback was realized by using an accelerometer instead of a microphone for onset detection of in-air tapping. In each combination of haptic/virtual condition and auditory

³musical instrument digital interface (MIDI): <https://midi.org/>

tapping strength, participants were asked to tap at constant rhythm (1 Hz) for 80 s. After each block, participants had to fill out a large number of questionnaires on emotions that are rather irrelevant in this context, but also a Likert-item on perceived surface hardness. As expected, the perceived hardness of the physical plate was significantly higher than that of the virtual plate. In addition, the virtual surface was perceived significantly harder for strong auditory feedback than for weak auditory feedback.

Lederman (1979) let participants judge the roughness of surface texture in a magnitude-estimation task under three conditions: unisensory auditory, unisensory haptic, and multisensory auditory-haptic. The true roughness was controlled by the width of grooves in a physical metal plate. In all three conditions, an increase in true roughness led to increased perceived roughness. Interestingly, estimated roughness was almost identical in the haptic and auditory-haptic conditions, while it was significantly different in the unisensory auditory condition. In particular, for small to moderate true roughness, auditory judgments were significantly higher than in the presence of haptic feedback, while for high true roughness, estimations were almost identical between all conditions. The results suggest that for haptic properties, congruent but redundant information from audition is simply ignored in the presence of haptic cues. In the lack of haptic feedback, however, auditory feedback is capable of predicting haptic information to a large extent.

Fernström and Brazil (2004) simulated buttons on a touchscreen device without visual feedback only by the use of sound. An arrangement of buttons could be explored by moving the fingers across the surface. Auditory feedback for buttons was synthesized by a simple friction model; boundaries were audible by an impact model upon entry and exit of the active area of a button. In informal evaluations, participants were able to draw the correct button layout. The authors suggest that auditory feedback allows users to have a pseudo-haptic experience which supports the development of a mental model of the button layout.

References

Fernström, Mikael and Eoin Brazil (2004). "Human-Computer Interaction Design based on Interactive Sonification – Hearing Actions or Instruments/Agents." In: *Interactive Sonification Workshop (ISon)*. Bielefeld, Germany.

Furfaro, Enrico et al. (2013). "Sonification of surface tapping: influences on behavior, emotion and surface perception." In: *Interactive Sonification Workshop (ISon)*. Erlangen, Germany.

Furfaro, Enrico et al. (2015). "Sonification of virtual and real surface tapping: evaluation of behavior changes, surface perception and emotional indices". In: *IEEE MultiMedia*, pp. 1–1. DOI: 10.1109/MMUL.2015.30.

Kjøer, Hans Peter, Christian Clive Taylor, and Stefania Serafin (2007). "Influence of interactive auditory feedback on the haptic perception of virtual objects." In: *Interactive Sonification Workshop (ISon)*. York, UK.

Lederman, Susan J (1979). "Auditory texture perception." In: *Perception* 8, pp. 93–103.

McGee, Marilyn Rose, Philip D. Gray, and Stephen Brewster (2002). "Mixed Feelings: Multimodal Perception of Virtual Roughness". In: *Proceedings of Eurohaptics*, pp. 47–52.

McGee, Marilyn Rose, Philip D. Gray, and Stephen A. Brewster (2001). "The effective combination of haptic and auditory textural information". In: *Lecture Notes in Computer Science*. Vol. 2058. Springer Berlin Heidelberg, pp. 118–126. DOI: 10.1007/3-540-44589-7_13.

Serafin, Stefania et al. (2007). "Audio-Haptic Physically Based Simulation and Perception of Contact Textures." In: *International Conference on Auditory Display (ICAD)*. Montréal, Canada, pp. 203–207.

2.5.2.4. Surface texture

The fine structural details of surfaces are generally referred to as texture, corresponding to a multidimensional percept that includes information from different sensory modalities such as audition, vision, and touch. Multisensory texture perception has been thoroughly reviewed by Klatzky and Lederman (2010) (also Lederman and Klatzky 2004). In general, texture perception seems to be mainly driven by tactile sensations. While auditory cues conveyed congruent information on surface texture and led to similar roughness judgments as tactile cues in case of unisensory perception, in presence of both sensory modalities, participants usually relied uniquely on tactile sensations.

References

- Klatzky, Roberta L. and Susan J. Lederman (2010). "Multisensory Texture Perception". In: *Multisensory Object Perception in the Primate Brain*. Ed. by J. Kaiser and M. J. Naumer. New York, NY: Springer, pp. 211–230. DOI: 10.1007/978-1-4419-5615-6_12.
- Lederman, Susan J. and Roberta L. Klatzky (2004). "Multisensory texture perception." In: *The Handbook of Multisensory Processes*. MIT Press, pp. 107–122.

2.5.2.5. Subjective quality

Altinsoy and Altinsoy (2012) evaluated the effect of auditory, haptic, and auditory-haptic feedback due to knocking interaction on the subjective quality of household appliances. A first experiment used only auditory stimuli including recordings of impacted washing machines and synthesized sounds of impacted plates. Participants rated these based on a semantic differential scale including 12 attributes (overall quality, pleasant, solid, loud, high-frequency, muffled, hard, rickety, lingering, strong, tonal, expressive). A higher overall quality was connected to strong low frequencies and weak high frequencies, sharp but quiet sounds instead of sharp and loud sounds. Low perceived quality was connected to very loud and very quiet sounds, or particularly high levels in any other attribute. A factor analysis revealed three combined factors that explained 79% of the variance. They were labeled quality, pleasantness, and solidity. Based on these results, the authors developed a model that predicts the quality index on the basis of a knocking stimulus via acoustical descriptors such as sharpness, tonality, loudness, and decay time. The model achieved a Pearson correlation coefficient of 0.89 for predicting subjective quality.

In a second experiment, blindfolded participants themselves knocked on the sidewalls of the washing machines either without (auditory-haptic) or with noise-blocking headphones (haptic). In about half of the stimuli, the sensory modality had no significant effect on perceived overall quality. In some cases, the subjective quality in the multisensory condition equaled the average of the other conditions, in other cases, it equaled the result of one dominant modality, sometimes haptic, sometimes auditory.

References

- Altinsoy, M. Ercan and M. Ercan Altinsoy (2012). "Knocking Sound as Quality Sign for Household Appliances and the Evaluation of the Audio-Haptic Interaction". In: *Haptic and Audio Interaction Design (HAID)*. Ed. by Charlotte Magnusson, Delphine Szymczak, and Stephen Brewster. Red. by Takeo Kanade et al. Vol. 7468. Springer Berlin Heidelberg, pp. 121–130. DOI: 10.1007/978-3-642-32796-4_13.

2.5.2.6. Summary

Auditory and haptic feedback generally originate from the same underlying physical process: mechanical vibration. Their perception is therefore tightly connected. We rely on their combined information in many contexts, e.g., if vision is occupied by other tasks such as during walking. Our walking style is therefore manipulated by sound parameters such as spectral centroid or roughness of the auditory feedback, and even by the overall soundscape. In ground material identification, there is even a perceptual dominance of audition over haptics, between 60% and 100%. In general, auditory feedback dominates for impacts, whereas tactile feedback dominates in case of incongruent sensory information.

Perception of surface texture or roughness is generally dominated by haptics (auditory feedback is often even ignored in the presence of haptic feedback), but auditory feedback is capable of altering the percept. It may even predict haptic properties and thus serve as a substitute to create a pseudo-haptic experience if no haptic feedback is present.

Estimations of the quality of products via knocking can be predicted by auditory cues alone.

2.5.3. Auditory-visual feedback

2.5.3.1. Material category

Fujisaki et al. (2014) examined sensory integration between vision and sound based on auditory-visual stimuli constructed from computer-generated videos in conjunction with recordings of real impact sounds. Visual stimuli showed a human hand hitting a rigid object with a stick; in 6 different variations of object texture corresponding to glass, ceramics, metal, stone, wood, and bark. Auditory stimuli were impact sounds of real objects made of glass, ceramic, metal, stone, wood, vegetable (pepper), plastic, and paper, hit with a wooden mallet. Stimuli were extended

by a blank visual and silent audio stimulus for audio-only and video-only conditions, respectively. All possible combinations of auditory and visual stimuli were displayed to participants of the experiment. Participants rated plausibility on a scale between “cannot be thought of” and “can be thought of” for each item in a list of 13 materials (the 8 already mentioned, extended by vinyl, cloth, clay, and leather). Strong auditory-visual interactions in material category perception were found. Some combinations led to an auditory-induced change of the perceived visual material category; e.g., visual glass and auditory vegetable in combination perceived as plastic. Other combinations led to a vision-induced change of the perceived auditory material category; e.g., auditory wood and visual glass in combination perceived as plastic. Multisensory material perception could be predicted by a multiplicative model from the unisensory results, indicating an AND-like sensory integration of information from the two sensory modalities. This model does not hold for implausible combinations; in such cases, a regression analysis indicated a stronger weight of auditory cues. This dominance of audition is explained by experience from everyday life that the true underlying material may be visually hidden by a surface coating that may be deceptive. Fujisaki et al. infer that the multiplicative integration rule may be universal to multisensory integration for categorical judgments.

References

Fujisaki, W. et al. (Apr. 17, 2014). “Audiovisual integration in the human perception of materials”. In: *Journal of Vision* 14.4, pp. 12–12. DOI: 10.1167/14.4.12.

2.5.3.2. Veracity and plausibility

Bonneel et al. (2010) examined sensory interaction between auditory and visual information on perceived material properties of rigid-body objects in virtual environments. Stimuli were combinations of real-time renderings of two shapes (bunny and dragon) and two materials (gold and plastic). Auditory and visual stimuli were rendered with variable level of detail (LOD). Visual LOD was controlled in the spherical harmonics domain by adjusting the number of coefficients used to render light reflectance. Auditory LOD was controlled by the number of rendered partials of a modal synthesis model. Both approaches refer to a projection of a scalar field into a set of functional bases and are

thus assumed to consistently provide a similar type of reconstruction error. A stimulus consisted of a sequence of the object falling and bouncing on a rigid surface. In pairwise comparisons, gave similarity ratings, while one stimulus was the reference with highest possible LOD, the other stimulus varied in auditory and/or visual LOD. Similarity ratings were assumed to indirectly express the quality of the rendering. The extreme values of LOD were displayed to the participants prior the experiment to define the range of the similarity scale. While higher visual LOD and higher auditory LOD generally led to higher similarity ratings, a statistical interaction between vision and sound was found. The same similarity could be achieved by high visual quality and low auditory quality, as well as by the other way around. The authors argue that high quality sound is capable of counteracting a low visual quality and vice versa.

Alaerts et al. (2009) used transcranial magnetic stimulation (TMS) to measure participants’ responses of the primary motor cortex (M1) by means of surface electromyogram (EMG). Participants were confronted with unisensory auditory (A) or visual (V) or multisensory (AV) stimuli (videos) of simple hand actions while corticomotor excitability in M1 was measured. Participants were asked to reproduce the performed action simultaneously. The 6 different stimuli showed a hand crushing a plastic bottle in A, V, and AV condition, two incongruent versions (crushing video with the sound of floating water, crushing sound with the video of foot interaction), as well as a silent white background video as baseline condition. The response to TMS was similar for the unisensory stimuli while their sum equaled the response to the congruent AV stimulus. The responses to the incongruent AV stimuli were significantly lower than response to the congruent AV stimulus. The authors argue for a shared modality-dependent action representation at the level of the primary motor cortex. These representations are assumed to be evoked more robustly in case of congruent auditory and visual stimuli. Alaerts et al. suggest that multisensory perception is a feature of the mirror neuron system and applies not only to speech perception but also to action perception in general.

Bonebright (2012) examined the effect of visual context on the perceived veracity of everyday sounds. They recorded scenes of interaction with everyday objects (including paper ripping, spoon dropping, book shutting, etc.) as well as corresponding similar sounds (e.g., balloon popping instead of book shut-

2. The psychophysics of auditory feedback: an annotated bibliography

ting) and contrasting sounds (e.g., ratchet turning instead of spoon dropping). Similar and contrasting combinations of video and audio were produced by standard sound editing methods. For each stimulus, participants were asked to freely identify the sound, and to give a confidence rating as well as a rating of the veracity of the sound. While participants achieved almost perfect sound identification for the actual sounds (95% correct answers), contrasting sounds were often wrong (61%) and similar sounds were almost never correctly identified (14%). This pattern could also be confirmed by the confidence ratings. Participants were thus greatly influenced by the visual action even if told to identify the sound alone. Veracity ratings were significantly higher for actual than for similar sounds and significantly higher for similar than for contrasting sounds. This was in conflict with the expectations that similar sounds would be perceived as perfectly real as is usually assumed in sound design for cinema.

Götzen et al. (2013) performed a similar experiment with videos showing a man walking on different ground materials (wooden pavement, concrete, sand, gravel, grass). For each video, real, Foley, and synthetic sounds were created. Real sounds were recordings matching the individual ground material, taken from the internet. Foley sounds were recorded in a Foley pit by standard Foley techniques. The synthetic sound was derived from the original sound by using a linear prediction filter with 48 coefficients on a white noise input signal, in combination with the envelope of the original sound. Two kinds of environmental soundscapes (outdoor, apartment) were added to compensate the cleanliness of the synthesized and Foley sounds. In two conditions (auditory, auditory-visual), naive participants were asked to describe shown action, as well as to rate the amount of realism. In case of auditory presentation, actions were generally better recognized than materials; wood, concrete, and gravel had better identification rates than grass and sand; participants had difficulties in the recognition of synthesized sounds. In case of bimodal auditory-visual presentation, realism was significantly higher for real and Foley than for synthesized sound, and higher for gravel, concrete, and wood than for sand and grass.

References

Alaerts, Kaat, Stephan P. Swinnen, and Nicole Wenderoth (Oct. 2009). "Interaction of sound and sight during action perception: Evidence

for shared modality-dependent action representations". In: *Neuropsychologia* 47.12, pp. 2593–2599. DOI: 10.1016/j.neuropsychologia.2009.05.006.

Bonebright, Terri L (2012). "Were those coconuts or horse hoofs? Visual context effects on identification and perceived veracity of everyday sounds". In: *International Conference on Auditory Display (ICAD)*. Atlanta, GA.

Bonneel, Nicolas et al. (Jan. 1, 2010). "Bimodal perception of audio-visual material properties for virtual environments". In: *ACM Transactions on Applied Perception* 7.1, pp. 1–16. DOI: 10.1145/1658349.1658350.

Götzen, Amalia de et al. (2013). "Real, Foley or synthetic? An evaluation of everyday walking sounds". In: *Sound and Music Computing Conference (SMC)*. Stockholm, Sweden, pp. 487–492.

2.5.3.3. Influence on grasping actions

Castiello et al. (2010) examined the effect of contact sounds on grasping interaction. Participants were sitting in front of a 8 cm plastic sphere packed in one of 4 materials (aluminum, paper, string, wool) with the task to reach out to grasp it. Pre-recorded contact sounds of the same objects were played either before (at start of the trial) or following the reaching-to-grasp movement (triggered at a certain distance before actual grasping). Participants' grasping behavior was measured under 3 conditions: (1) congruent, i.e., matching auditory and visual cues, (2) incongruent, i.e., with the sound of a different material, and (3) control, with an unrelated synthetic sound. Reaction time, movement duration, and grip closing time were shorter for congruent information compared to the control condition, and longer in case of incongruent information. Congruent information thus facilitated the task while incongruent information led to sensory interference. In a second experiment, the upper and lower parts of the stimulus objects were covered with different materials, and the presented sound was always congruent to one of the two materials. In result, participants consistently (in 84% of the trials) grasped the object at the material that was congruent to the sound. The results support the assumption that contact sounds are tightly integrated in the feedback loop of the planning system, as they provide critical information on physical object properties such as fragility, texture, and weight.

In a similar type of experiment, Sedda et al. (2011) asked participants to blindly reach out and

grasp a target object while grip aperture was measured. The experiment was performed in 4 combinations of vision/blind and sound/silence, with two different target sizes (small, big). In case of vision, participants were able to see the target for 200 ms before starting their grasping movement. In case of sound, the experimenter placed the object at the start of the trial in a way that an audible collision sound was produced. In general, grip aperture was significantly larger for big than for small target size, except for the blind/silent condition. In the silent condition, the reaction times were significantly shorter with vision than blind. Wrist velocity was faster with vision; both auditory and visual cues increased wrist velocity for the big stimulus object. In summary, consistent with the previously described experiment, auditory information about the object size prior to a goal-oriented action was shown to improve the execution of the action.

References

- Castiello, Umberto et al. (Aug. 17, 2010). "When Ears Drive Hands: The Influence of Contact Sound on Reaching to Grasp". In: *PLoS ONE* 5.8, e12240. DOI: 10.1371/journal.pone.0012240.
- Sedda, Anna et al. (Mar. 2011). "Integration of visual and auditory information for hand actions: preliminary evidence for the contribution of natural sounds to grasping". In: *Experimental Brain Research* 209.3, pp. 365–374. DOI: 10.1007/s00221-011-2559-5.

2.5.3.4. Summary

From cinema we know that vision is sometimes capable of influencing auditory perception, and vice versa. We are strongly influenced by the visual action, even when asked to judge sound alone. The widespread practice of substituting the original physical sound by Foley sounds, i.e., recordings that mimic only the relevant sound parameters but not the true physical process, shows us that we are generally open to alternative auditory feedback. For auditory-visual stimuli, we tend to identify sounds based on vision. However, if asked explicitly, we are capable of exposing auditory-visual discrepancies. A high sound quality can counteract low visual quality and vice versa.

The unisensory information from vision and audition is merged in a multiplicative manner, i.e., by AND-like sensory integration, at least for categorical judgments such as material category. This means

that a multisensory stimulus is identified as a specific material only, if no single unisensory feedback is in conflict to that result. In general, multisensory perception seems to be a feature of the mirror neuron system. Congruent auditory-visual information facilitates certain tasks (e.g., leading to shorter reaction time or movement duration) while incongruent information leads to sensory interference and thus deteriorated task performance. The fact that unrelated contact sounds even influence a future grasping action suggests that these are tightly integrated in the feedback loop of the planning system.

2.5.4. Auditory-visual-haptic feedback

2.5.4.1. Material properties

Martín et al. (2015) examined the multisensory perception of material properties of 12 different materials including leathers, papers, and fabrics. Participants were asked to rate material properties in different sensory conditions: auditory (playback of pre-recorded contact sounds), visual (an image of the material surface), auditory-visual (a combination of both), and full-modal (interaction with a physical sample of the material). 10 material properties were rated on a continuous scale defined by pairs of opposite properties. These affected tactile (rough/smooth, hard/soft, warm/cold), visual (shiny/matte, simple/complex, colorful/colorless), and subjective (expensive/cheap, old/new, natural/synthetic, beautiful/ugly) dimensions. It was assumed that a property is clearly transported for a certain stimulus if there is strong agreement, i.e., inter-participant correlation (IPC), between participants for this quality. On the other hand, if the stimulus doesn't convey information about a certain property, participants are assumed to rely on their imagination, leading to reduced IPC. Furthermore, if a certain condition is well suited for the representation of a certain property, participants' ratings are assumed to be close to the full-modal condition. If ratings are far apart from the full-modal condition, the condition is assumed to be less realistic. Results showed that participants generally agreed on tactile properties for auditory stimuli, while for visual stimuli they showed high IPC for visual properties. Subjective properties led to significantly higher IPC for visual than for auditory representation. The results of the bimodal auditory-visual condition suggested an additive type of sensory integration, i.e., auditory-visual ratings roughly equaled the sum of auditory and visual ratings. A principal component

2. *The psychophysics of auditory feedback: an annotated bibliography*

analysis (PCA) revealed the number of significant dimensions of the perceptual property space: 1 for auditory, 2 for visual, 3 for auditory-visual, 4 for full-modal.

A second experiment by Martín et al. investigated cross-modal effects between vision and sound. Participants were asked to rate auditory, visual, and auditory-visual representations based on a subset of the previous properties (all tactile properties, plus old/new, beautiful/ugly, and unrealistic/believable). In addition to congruent representations, all possible combinations of auditory and visual stimuli were tested. Overall, varying visual information did not significantly influence the perception of tactile properties, indicating a dominance of auditory perception.

Martín et al. conclude that sound offers an orthogonal complement to the visual channel, and that its addition benefits the perception of digital materials, especially for tactile qualities. According to the authors, sound is capable of deliberately inducing a bias which manipulates the perception of those qualities.

In a follow-up experiment (Martín et al. 2018), an additional active auditory-visual condition was added to the passive one, implemented via real-time granular sound synthesis of rubbing sounds connected to a touchscreen. In this case, the addition of rubbing sounds did not lead to a significant improvement in perception of material qualities. This was attributed to the limitation to inferior sound synthesis that neglected important aspects such as the impact sound on first contact. However, the presence of interactive audio raised the level of immersion.

Fujisaki et al. (2015) used a similar procedure to evaluate the perception of material properties of 22 different specimen of wood (including genuine, processed, and fake laminated types) under 3 unisensory conditions: auditory (recorded mallet impact), visual (photograph), and tactile (blind exploration with the index finger). Participants gave ratings on 23 scales defined by bipolar adjective pairs, including 3 visual, 3 auditory, 6 tactile/other, and 11 affective or preferential properties.

Results showed that evaluations of affective properties were similar across sensory modalities. A factorial analysis suggested that affective properties of wood are at least partly represented in a supramodal fashion. Consistent over all 3 senses, affective properties split in 2 distinct groups (expressiveness, sturdiness, rareness, interestingness, and sophisticatedness vs. pleasantness, relaxed feel-

ings, and liked/disliked). IPC were generally high for surface brightness, sound sharpness and surface roughness; however, a subsidiary experiment comparing stone and wood suggested that this pattern differs across materials.

By re-analyzing the gathered data from the previous experiment, Kanaya et al. (2016) inferred cross-modal correspondences between vision, audition, and touch, with a focus on the properties surface brightness, sharpness of sound, and surface smoothness. For each property, pairwise correlations were performed between the ratings from the individual sensory modalities. For brightness, a significant positive correlation between audition and touch suggests commonalities between those two modalities. Both, however, showed negative correlation with vision, showing that participants were actually unable to correctly judge this visual property by modalities other than vision. Brightness was instead predicted by auditory dampenedness and dullness and tactile roughness, warmth, and dryness. Participants were unable to judge sharpness of sound by vision while trying to predict it from glossiness and darkness. Tactile perception of sharpness of sound was instead successfully predicted by denseness and coldness. If judging smoothness of touch, participants were able to successfully predict the tactile impression from visual cues (glossiness, clearness, darkness) or auditory cues (sharpness, dampenedness).

Avanzini and Crosato (2006b) simulated rigid objects of different stiffness in an interactive setup where participants interacted by means of a Phantom Omni haptic device, a visual 3D rendering projected on a 2D computer screen, and real-time sound synthesis presented via headphones (see also Avanzini and Crosato 2006a). Naive participants without prior training were asked to judge the stiffness of the impact between the hammer and a bar clamped on one end while interacting freely within the virtual environment. Absolute magnitude estimations were given on an ordinal scale ranging from extremely soft to extremely stiff. Average perceived stiffness generally increased with higher true stiffness, with saturating effects at very low and very high true stiffness. In a post-hoc interview, 40% of the participants told that they based their judgments on both auditory and haptic feedback. The authors conclude that sound is indeed capable of modulating haptic perception.

References

- Avanzini, Federico and Paolo Crosato (2006a). "Haptic-Auditory Rendering and Perception of Contact Stiffness". In: *Haptic and Audio Interaction Design (HAID)*. Vol. 4129. Springer Berlin Heidelberg, pp. 24–35. DOI: 10.1007/11821731_3.
- (July 2006b). "Integrating physically based sound models in a multimodal rendering architecture". In: *Computer Animation and Virtual Worlds* 17.3, pp. 411–419. DOI: 10.1002/cav.144.
- Fujisaki, Waka, Midori Tokita, and Kenji Kariya (Apr. 2015). "Perception of the material properties of wood based on vision, audition, and touch". In: *Vision Research* 109, pp. 185–200. DOI: 10.1016/j.visres.2014.11.020.
- Kanaya, Shoko, Kenji Kariya, and Waka Fujisaki (Oct. 2016). "Cross-Modal Correspondence Among Vision, Audition, and Touch in Natural Objects: An Investigation of the Perceptual Properties of Wood". In: *Perception* 45.10, pp. 1099–1114. DOI: 10.1177/0301006616652018.
- Martín, Rodrigo, Michael Weinmann, and Matthias B. Hullin (2018). "A Study of Material Sonification in Touchscreen Devices". In: *International Conference on Interactive Surfaces and Spaces*. Tokyo, Japan: ACM Press, pp. 305–310. DOI: 10.1145/3279778.3281455.
- Martín, Rodrigo et al. (2015). "Multimodal perception of material properties". In: *ACM SIGGRAPH Symposium on Applied Perception*. Tübingen, Germany: ACM Press, pp. 33–40. DOI: 10.1145/2804408.2804420.

2.5.4.2. Naturalness, usability, and pleasantness

(Susini et al. 2012) investigated the influence of the naturalness of auditory feedback on the perceived usability and pleasantness of a graphical user interface using the example of an ATM interface. Three different sound mappings for keyboard interactions were created. A causal mapping of high naturalness was based on recordings taken from a sound library. An iconic mapping of medium naturalness was created via cross synthesis based on the causal mapping. An arbitrary and incongruent mapping of low naturalness consisted of random sounds (bicycle ring, piano chord, etc.) uncorrelated to the performed action. The mappings were explained to the participants via video examples of a different action. Participants then rated naturalness, usability,

and pleasantness of the sounds; first without context, then after interacting with the interface and directly comparing the different mappings. The same experiment was carried out in a normal as well as in a defective mode of operation. On average, a higher true naturalness led to a significantly higher subjective naturalness, usability, and pleasantness. A causal relationship resulted in sounds being perceived as natural, in contrast to arbitrary sounds perceived as unnatural. Natural sounds were perceptually more pleasant and more usable than unnatural sounds. In addition, the perceived naturalness and usability of the sounds increased after interacting with the interface.

References

- Susini, Patrick et al. (May 2012). "Naturalness influences the perceived usability and pleasantness of an interface's sonic feedback". In: *Journal on Multimodal User Interfaces* 5.3, pp. 175–186. DOI: 10.1007/s12193-011-0086-0.

2.5.4.3. Influence on typing performance

Ma, Zhaoyuan et al. (2015) examined the effect of haptic and auditory feedback on the typing performance with a computer keyboard. The experiment was performed with a flat keyboard without moving keys, but enabled with haptic feedback through piezo actuators under each key, driven by 3 cycles of a 250 Hz signal if triggered. Baseline full-modal performance was measured with a classic keyboard exhibiting mechanically-movable keys. Auditory feedback was provided by means of the standard key-click sound of the Microsoft Surface Pro. It must be noted that the authors measured auditory latency to equal 107 ms which is surely perceptible and assumed to be disturbing. The original auditory feedback was blocked by headphones playing pink noise. Participants were asked to typewrite a given text under different conditions including full-modal, visual, auditory-visual, visual-haptic, and auditory-visual-haptic. Overall typing speed was found to be significantly higher with increasing number of sensory modalities. Only in the presence of haptic feedback, auditory feedback had no significant effect on typing speed. The exact same pattern, but inverse, could be found in the rate of typing errors. The authors conclude that haptic feedback is generally preferred to auditory feedback for improving typing performance.

References

Ma, Zhaoyuan et al. (June 2015). "Haptic keyclick feedback improves typing speed and reduces typing errors on a flat keyboard". In: *World Haptics Conference (WHC)*. Evanston, IL: IEEE, pp. 220–227. DOI: 10.1109/WHC.2015.7177717.

2.5.4.4. Summary

A human–computer interface can reach high naturalness by incorporating natural, causal, and congruent feedback in the sensory modalities audition, vision, and touch. The result is not only more pleasant for the user, but offers even higher usability. While Audition and vision are generally regarded as orthogonal sensory channels, audition and touch often convey the same information redundantly. In many tasks such as typing, auditory feedback can partially replace haptic feedback, but adds only little benefit if provided together with haptic feedback. However, the addition of interactive sound usually increases the level of immersion.

The perception of individual material properties follows an additive law of sensory integration. The multisensory perception of a specific material property approximately equals the sum of its estimation in the individual sensory channels alone. How well a property is conveyed by a certain stimulus can be inferred from the inter-participant correlation (IPC) between individual participants for the specific property.

3. Physical modeling for plausible auditory augmentation



This chapter includes the same theoretic foundations and synthesis model as described by Czuka (2021) in his master's thesis under my supervision and in the corresponding conference paper (Czuka, Weger, and Höldrich 2021).

If sonifications should integrate into the physical world in a consistent manner, they should also sound like the physical world. We therefore need not only to understand how certain physical properties affect certain aspects of their sound, but also how to synthesize such sounds. And even if we were able to describe the underlying physical processes in great detail, we still need to implement them so that they can run in real time on an ordinary computer—or in an actual application on a small microcomputer that can be embedded into physical objects.

In general, we are still not able to cover all aspects of physical musical instruments or physical objects in a modal synthesis model. Electronic musical instruments are therefore still hardly reaching the level of embodiment and musical expression that traditional musical instruments provide. However, for certain types of interaction—especially in the “hit and listen” category—physical sound models have reached a degree of realism that is on par with the real physical sound. In Sec. 2.2.16 we already showed that humans are unable to distinguish between real and synthetic impact sounds, even with very simplified modal synthesis models (Lutfi et al. 2005; Traer et al. 2019). Such models have even been used extensively in psychoacoustic experiments on perception of physical parameters, under the assumption that the results remain their validity for real physical sounds (see Sec. 2.2).

This chapter is structured in three sections. First, in Sec. 3.1, we will examine the vibrational behavior of the very simplest vibrating objects. Section 3.2 then summarizes the acoustics of bars and plates in order to understand the relation between physical parameters and sound parameters. Finally, in Sec. 3.3, we will recapitulate the basics of modal analysis and modal synthesis and describe the physical sound model that forms the basis of experiments and sonifications described in the upcoming chapters.

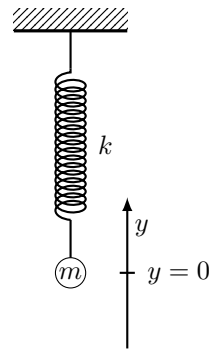


Figure 3.1.: An undamped harmonic oscillator consisting of a mass m and a spring k . The displacement of the mass is indicated by y .

3.1. Acoustics of simple systems

Modal analysis is an elegant way to describe the complex dynamic behavior of vibrating rigid objects with arbitrary shape on the basis of so-called modes. In modal synthesis, the contribution of each mode, oscillating with a single complex frequency, is synthesized separately, and the resulting individual vibrations are added to obtain the overall result. In order to fully understand this concept, we need to take a step back and examine the basic physical principles that form the foundations of this concept.

3.1.1. Undamped harmonic oscillator

Let us begin with the simplest vibrational systems: the undamped harmonic oscillator, consisting of a mass m that is held by a spring with stiffness k , as shown in Fig. 3.1. If gravity is neglected, only one force acts on the spring: the force $F_k = -ky$ of the spring which always pulls the mass towards the equilibrium displacement $y=0$. The force of a moving mass can also be described by $F = m \cdot a$, with acceleration a simply being the second derivative of displacement y with respect to time. A balance of forces $F = F_k$ leads to the ordinary differential equation

$$m\ddot{y} + ky = 0 \quad . \tag{3.1}$$

3. Physical modeling for plausible auditory augmentation

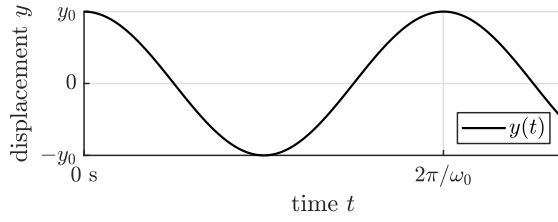


Figure 3.2.: The vibration of an undamped harmonic oscillator over time.

What we seek is the vibration of the system, i.e., the displacement $y(t)$ as a function of time t . A solution of Eq. 3.1 with respect to y , the so-called harmonic solution, is given by the function

$$y(t) = y_0 \sin(\omega_0 t + \phi_0) \quad (3.2)$$

with the angular natural frequency ω_0 . The displacement y_0 and phase angle ϕ_0 describe the initial conditions of the system, accounting for the displacement and phase offset of the sinusoid at time $t=0$. The resulting displacement over time, i.e., its vibration is shown in Fig. 3.2. To derive the value of ω_0 , the solution is inserted back into the differential equation, leading to the quadratic equation

$$-\omega_0^2 m + k = 0 \quad (3.3)$$

As physical systems are always causal, only the positive solution with respect to ω_0 ,

$$\omega_0 = \sqrt{\frac{k}{m}} \quad (3.4)$$

is considered. Via the relationship $\omega = 2\pi f$, it defines the frequency of the sine-shaped movement by using only the material constants mass m and stiffness k .

3.1.2. Damped harmonic oscillator

If an additional damper is added to the spring, as shown in Fig. 3.3, there is a second force acting on the mass: the damping force $F_r = -rv$. With velocity v being the 1st derivative of displacement y with respect to time, a balance of forces $F = F_k + F_r$ now leads to the extended differential equation

$$m\ddot{y} + r\dot{y} + ky = 0 \quad (3.5)$$

with its solution with respect to y

$$x(t) = y_0 e^{-\alpha t} \sin(\omega_d t + \phi_0) \quad (3.6)$$

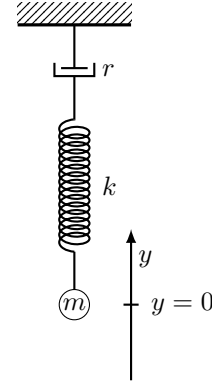


Figure 3.3.: A damped harmonic oscillator consisting of a mass m , spring k , and damper r . The displacement of the mass is indicated by y .

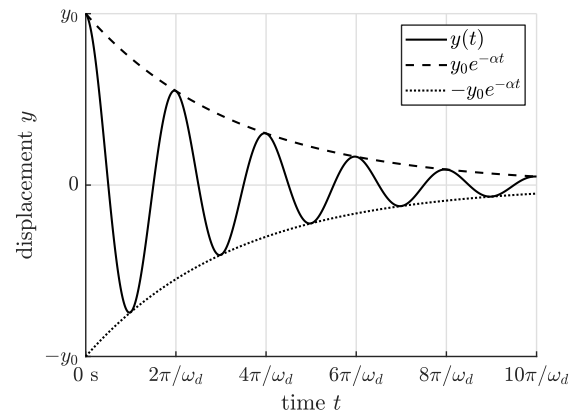


Figure 3.4.: The vibration of a damped harmonic oscillator over time.

The term $e^{-\alpha t}$ describes an envelope in the form of an exponential decay, with decay factor α being

$$\alpha = \frac{r}{2m} \quad (3.7)$$

Depending on the damping ratio ζ , with

$$\zeta = \frac{\alpha}{\omega_0} = \frac{r}{2\sqrt{mk}} \quad (3.8)$$

a damped system can be *overdamped* ($\zeta > 1$) where it just exponentially decays to the equilibrium without oscillation, *critically damped* ($\zeta = 1$) where it returns to the equilibrium as quickly as possible, or *underdamped* ($\zeta < 1$) where it oscillates with frequency ω_d while the amplitude exponentially returns to the equilibrium, as shown in Fig. 3.4.

Note that the frequency ω_d of the damped harmonic oscillator is slightly below that of the un-

damped system ω_0 :

$$\omega_d = \sqrt{\omega_0^2 - \alpha^2} = \omega_0 \sqrt{1 - \zeta^2} . \quad (3.9)$$

3.1.3. Driven harmonic oscillator

If the damped harmonic oscillator is subject to a sinusoidal driving force, that force is again added to the balance of forces, leading to

$$F = F_k + F_r + F_d . \quad (3.10)$$

with $F_d = F_0 \sin(\omega t)$, amplitude F_0 and driving frequency ω . The differential equation thus becomes

$$m\ddot{y} + r\dot{y} + ky = F_0 \sin(\omega t) \quad (3.11)$$

with its steady-state solution

$$y(t) = \frac{F_0}{mZ_m\omega} \sin(\omega t + \phi_d) , \quad (3.12)$$

with the impedance

$$Z_m = \sqrt{(2\omega_0\zeta)^2 + \frac{1}{\omega^2}(\omega_0^2 - \omega^2)^2} \quad (3.13)$$

and the phase of the oscillation relative to the driving force

$$\phi_d = \arctan\left(\frac{2\omega\omega_0\zeta}{\omega^2 - \omega_0^2}\right) . \quad (3.14)$$

Figure 3.5 shows the frequency response of the system, which is the amplitude of the oscillation plotted against the frequency of the driving force. The maximum amplitude is achieved at frequency ω_r , with

$$\omega_r = \sqrt{\omega_0^2 - 2\alpha^2} = \omega_0 \sqrt{1 - 2\zeta^2} . \quad (3.15)$$

The amplitude A_0 at ω_0 equals the quality factor Q or simply Q-factor of the resonance:

$$Q = \frac{1}{2\zeta} . \quad (3.16)$$

The amplitude A_r at the peak is slightly above, with

$$A_r = Q \frac{\omega_0}{\omega_d} = \frac{1}{2\zeta\sqrt{1 - \zeta^2}} . \quad (3.17)$$

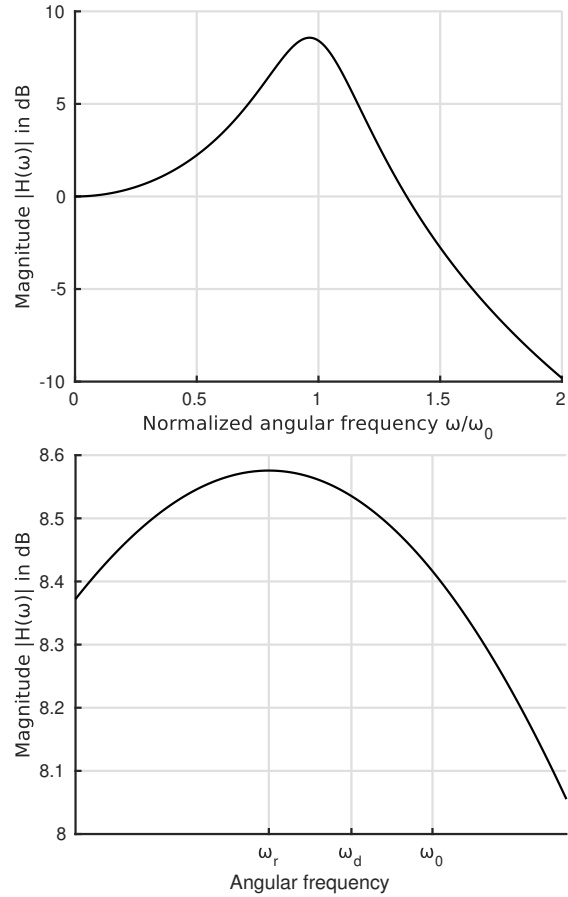


Figure 3.5.: Frequency response of a damped harmonic oscillator with physical parameters $m = 1$ kg, $k = 10^9$ kg s⁻¹, and $r = 10^3$ kg s⁻², leading to resonance at $f_0 = 5.033$ kHz with $Q = 2.635$.

3.1.4. Damping parameters and their relationships

We already defined the exponentially decaying envelope by $e^{-\alpha t}$, with decay factor α in Hz, as well as the non-dimensional damping ratio

$$\zeta = \frac{\alpha}{\omega} \quad (3.18)$$

which describes the quotient of actual damping and critical damping. Apart from those, the amount of damping and resonance can be expressed by a couple of other parameters which are more or less interchangeable.

The time constant τ is the time the system needs for dropping to the value $1/e$. Its value in seconds is simply:

$$\tau = \frac{1}{\alpha} . \quad (3.19)$$

3. Physical modeling for plausible auditory augmentation

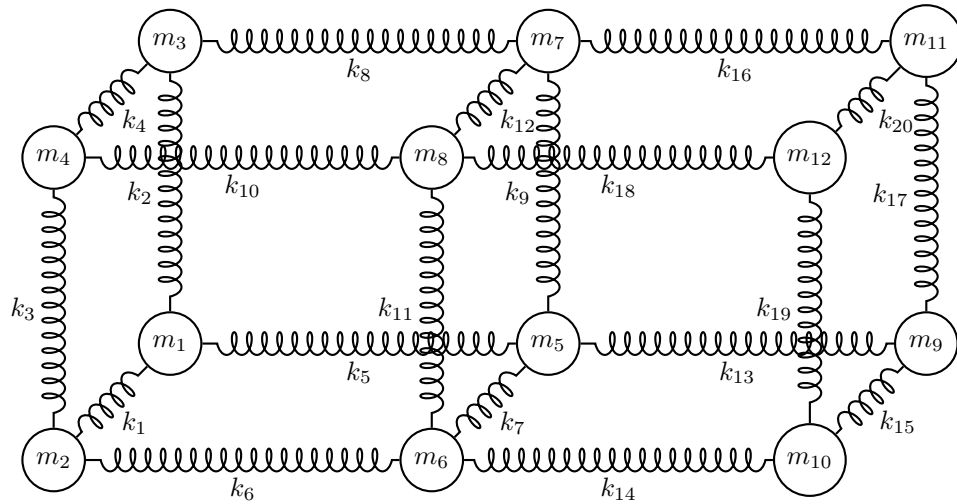


Figure 3.6.: Discretized model of a physical system, described by masses m_i and springs with stiffnesses k_j and dampers r_j . The dampers have been omitted in the drawing for clarity.

The -60 dB reverberation time T_{60} is the time the system needs for dropping to $1/1000$ of its original amplitude, and hence connects to τ via

$$T_{60} = \ln(1000)\tau = 6.9078\tau . \quad (3.20)$$

The loss factor relates to the other parameters via

$$\eta = 2\zeta = \frac{2\alpha}{\omega} = \frac{2}{\tau\omega} \quad (3.21)$$

and is equivalent to $\tan \delta$ which refers to the slope of the decay factor α (sometimes called δ). A constant loss factor leads to a decay factor $\alpha = \pi\eta f$ that rises linearly with frequency f , with slope $\pi\eta$.

The Q-factor is simply

$$Q = \frac{1}{\eta} . \quad (3.22)$$

The bandwidth B of a resonance equals the difference Δf between its upper and lower -3 dB cutoff frequency:

$$B = \Delta f = \frac{f_0}{Q} . \quad (3.23)$$

3.1.5. Coupled oscillators and modes

The vibrational behavior of rigid physical objects can be sufficiently described by a network of discrete masses which are connected through springs and dampers, as shown in Fig. 3.6. Due to the coupling springs, now the individual masses do not move independently anymore but form so-called modes.

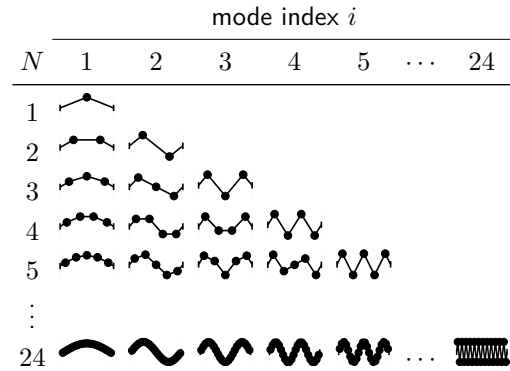


Figure 3.7.: The normal modes of a vibrating string for different numbers of masses N , leading to N modes. Adapted from Fletcher and Rossing (2010, p. 35).

Each state of the system can be expressed as a linear combination of these modes. Figure 3.7 shows the normal¹ modes of a vibrating string. Due to the discretization, the (otherwise infinite) number of modes equals the number of masses.

In real physical systems, the number of masses effectively approaches infinity, and the discrete system becomes a continuous system such as a string or bar, if extending in only one dimension, or a membrane or plate, if extending in two dimensions. Such systems are described in Sec. 3.2.

¹Normal modes consider only movement that is normal to the principal axis of the system or perpendicular to its surface.

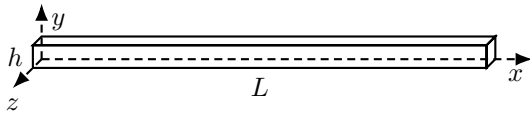


Figure 3.8.: A longish cuboid that is just thin enough to fulfill the requirements of Euler-Bernoulli beam theory for thin bars (relative dimensions: $20 \times 1 \times 1$).

3.2. Acoustics of continuous systems: bars and plates

In the previous section, complex dynamic systems were constructed from a finite number of coupled oscillators. Figure 3.7 already showed us that it might also be able to leave this discretization into individual masses behind us and instead assume a continuous distribution of infinite masses. For bars and rectangular plates, the governing differential equations can be solved analytically. This section does not only describe the computation of modes and their frequencies and shapes, but also the damping mechanisms and sound radiation of rectangular plates. It is intended to provide a mathematical conversion from physical parameters (size, shape, material constants, boundary conditions, initial conditions) to sound parameters (resonant frequencies, amplitudes, and quality factors) that may later drive a modal sound synthesis algorithm for (Sec. 3.3) for auditory augmentation.

3.2.1. Free vibrations of thin bars

Let us begin with the very simplest form of continuous vibrating objects—those that stretch out over just a single spatial dimension: thin bars of constant cross section. By thin, we actually mean “moderately thin”, which requires that the wavelength of the bending wave in the principal (length) dimension is at least $20 \times$ larger than all other dimensions (Steele and Balch 2009). This assumption allows the application of the “classical” Euler-Bernoulli beam theory. Such a bar is shown in Fig. 3.8. Within this framework, we only consider flexural vibrations normal to the longest dimension.

Under the assumption of an isotropic material, Hamilton’s principle lets us formulate the well-known equation

$$EI \frac{\partial^4 y}{\partial x^4} + \rho S \frac{\partial^2 y}{\partial t^2} = F(x, t) \quad (3.24)$$

with displacement y , longitudinal position x , second moment of inertia I , cross-sectional area S , Young’s

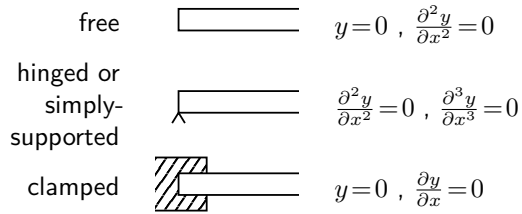


Figure 3.9.: Three basic end conditions of a bar (Fletcher and Rossing 2010, p. 61).

modulus E , density ρ , and an external driving force $F(x, t)$ (Fletcher and Rossing 2010, pp. 58–59). The displacement y depends on time t and position x and represents the transverse vibration of the bar. In the simplest case, there is no external force ($F(x, t) = 0$). Moment of inertia and cross-sectional area, both depending on the actual dimensions, are often combined to a so-called radius of gyration K , leading to the simplified form

$$\frac{\partial^2 y}{\partial t^2} + \frac{E}{\rho} K^2 \frac{\partial^4 y}{\partial x^4} = 0 \quad (3.25)$$

Fletcher and Rossing (2010, p. 59) provide different radii of gyration for some simple shapes:

$K = h/\sqrt{12}$ for rectangular cross section of thickness h , $K = R^2$ for a round cross section of radius R , and $K = \sqrt{R_1^2 + R_2^2}/2$ for a round pipe of inner radius R_1 and outer radius R_2 . The general harmonic solution to the differential equation yields

$$y(x, t) = \cos(\omega t) [A \cosh(kx) + B \sinh(kx) + C \cos(kx) + D \sin(kx)] \quad (3.26)$$

with the wave number $k = \omega/c$. In this context, c is the phase velocity of the flexural waves

$$c = \sqrt{\omega K c_L} \quad (3.27)$$

with the longitudinal wave velocity $c_L = \sqrt{E/\rho}$ (Rossing 2014, p. 961). The velocity of the flexural waves is proportional to $\sqrt{\omega}$. This frequency-dependency of velocity is called dispersion.

Since Eq. 3.26 is the solution of a 4th-order differential equation, we get four arbitrary real constants A , B , C , and D . To determine these, we need four boundary conditions: two for each end of the bar. The three most basic end conditions of a bar, in combination with their boundary conditions, are shown in Fig. 3.9.

3. Physical modeling for plausible auditory augmentation

As an example, we assume a bar that is free on both ends. The boundary conditions therefore dictate no bending moment and no shear force at the boundaries:

$$\left. \frac{\partial^2 y}{\partial x^2} \right|_{x=0,L} = \left. \frac{\partial^3 y}{\partial x^3} \right|_{x=0,L} = 0 . \quad (3.28)$$

If the solution (Eq. 3.26) is inserted into the boundary conditions (Eq. 3.28) at $x=0$, we get

$$C = A \text{ and } D = B . \quad (3.29)$$

These relations applied to the general solution, and altogether inserted in the two boundary conditions at $x=L$ yields the two equations

$$\begin{aligned} A [\cosh(kL) - \cos(kL)] \\ = -B [\sinh(kL) - \sin(kL)] \end{aligned} \quad (3.30)$$

$$\begin{aligned} A [\sinh(kL) + \sin(kL)] \\ = -B [\cosh(kL) - \cos(kL)] \end{aligned} \quad (3.31)$$

which have non-trivial common solutions only for certain values of kL and thus ω . Dividing the first equation by the second yields

$$1 - \cosh(kL) \cos(kL) = 0 \quad (3.32)$$

whose zeros define the allowed values of kL . Equation 3.32 can be numerically solved. The solutions approximate to

$$k_n L = \frac{\pi}{2} \{3.011, 5, 7, 11, \dots\} \quad (3.33)$$

and are inserted in Eq. 3.27 to yield angular frequencies ω_n and thus frequencies f_n :

$$f_n = \frac{\pi K}{8L^2} \sqrt{\frac{E}{\rho}} \{3.011^2, 5^2, 7^2, \dots, (2n+1)^2\} . \quad (3.34)$$

The general solution of the vibration becomes

$$y(x, t) = A_n \cos(\omega_n t + \phi_n) \Theta_n(x) \quad (3.35)$$

where $\Theta_n(x)$ is the characteristic beam function representing the spatial mode shape:

$$\begin{aligned} \Theta_n(x) = \cos(k_n x) - \gamma_n \sin(k_n x) \\ + \cosh(k_n x) [1 - \gamma_n \tanh(k_n x)] \end{aligned} \quad (3.36)$$

with

$$\gamma_n = \frac{\cos(k_n L) - \cosh(k_n L)}{\sin(k_n L) - \sinh(k_n L)} . \quad (3.37)$$

The first mode shapes of the free vibrating bar are shown in Fig. 3.10. The first two modes refer to translation and rotation, and have a frequency of zero.

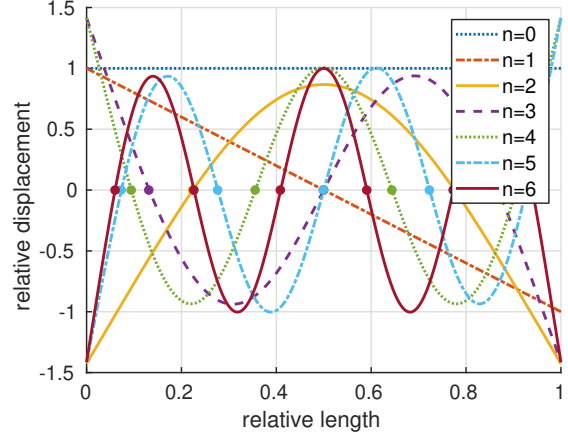


Figure 3.10.: The mode shapes of the first 7 flexural modes of a thin bar with free boundary conditions (including translation and rotation). n indicates the number of nodal lines, i.e., zero-crossings (\cdot).

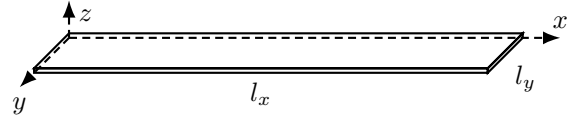


Figure 3.11.: A longish cuboid that is just thin enough and wide enough to fulfill the requirements of a thin plate (relative dimensions: $100 \times 20 \times 1$).

3.2.2. Free vibrations of thin plates

In case of a thin plate (see Fig. 3.11), also described as Kirchhoff-Love plate, we apply a generalization of the Euler-Bernoulli beam theory in 2D. We denote the longest side as length l_x and the shortest side as width l_y . Thickness h is assumed to be very small in comparison to length and width. In analogy to a plane being defined by two vectors, we can think of the plate being span by two bars. We assume that these bars are homogeneous themselves, but their elastic material constants might differ between the plate's two principal axes. This special case of anisotropy is called orthotropy. We therefore speak of a homogeneous orthotropic plate. Assuming that the plate lies in the x/y -plane of a Cartesian coordinate system, we denote $w(x, y, t)$ as the plate's transverse displacement. Similar to the bar, an equation of motion can be derived using

Table 3.1.: Basic material constants of isotropic and orthotropic plates.

constant	description	units
$\sigma / \sigma_{xy}, \sigma_{yx}$	Poisson's ratio (isotropic / orthotropic)	–
$E / E_x, E_y$	Young's modulus (isotropic / orthotropic)	Pa = kg m ⁻¹ s ⁻²
ρ	density	kg m ⁻³
G_{xy}	in-plane shear modulus	Pa = kg m ⁻¹ s ⁻²
D / D_1-D_4	rigidity (isotropic / orthotropic)	Pa = kg m ⁻¹ s ⁻²
Ω	orthotropy factor	–
HB	Brinell hardness	kgf = 9.806 65 N

Hamilton's principle (Rossing 2014, p. 964):

$$\rho \frac{\rho}{h^2} \frac{\partial^2 w}{\partial t^2} + D_1 \frac{\partial^4 w}{\partial x^4} + D_3 \frac{\partial^2}{\partial y^4} + (D_2 + D_4) \frac{\partial^4 w}{\partial x^2 \partial y^2} = 0 \quad (3.38)$$

If the plate is cut in parallel to one of the principal planes of the orthotropic solid, the four rigidity constants D_i can be expressed by more familiar elastic constants (McIntyre and Woodhouse 1988): Young's modulus E_x in x -direction and E_y in y -direction, and in-plane Poisson's ratios ν_{xy} and ν_{yx} , as well as in-plane shear modulus G_{xy} . The conversions are:

$$D_1 = \frac{E_x}{12\mu}, \quad D_2 = \frac{\nu_{xy} E_y}{6\mu} = \frac{\nu_{yx} E_x}{6\mu}, \quad (3.39)$$

$$D_3 = \frac{E_y}{12\mu}, \quad \text{and} \quad D_4 = \frac{G_{xy}}{3},$$

with $\mu = 1 - \nu_{xy}\nu_{yx}$.

The described elastic constants only cover isotropic materials such as glass or orthotropic materials such as wood. With wood, we always refer to quarter-cut lumber that is cut parallel to the fiber (grain). We now introduce an orthotropy factor

$$\Omega = \left(\frac{D_1}{D_3} \right)^{1/4} = \left(\frac{E_x}{E_y} \right)^{1/4} = \left(\frac{\nu_{xy}}{\nu_{yx}} \right)^{1/4}. \quad (3.40)$$

together with isotropic material constants

$$E = \sqrt{E_x E_y}, \quad \nu = \sqrt{\nu_{xy} \nu_{yx}}, \quad (3.41)$$

$$D = \sqrt{D_1 D_3} = \frac{E}{12(1 - \nu^2)}.$$

The shear modulus actually cannot be computed from Young's modulus and Poisson's ratio in the orthotropic case. Morozov and Vasiliev (2003) failed in deriving it from E_x , E_y , ν_{xy} , and ν_{yx} . Based on

their data, we found that the following approximation is sufficient for our needs:

$$G_{xy} \approx \min \left\{ \Omega^2, \frac{1}{\Omega^2} \right\} \frac{E}{2(1 + \nu)}. \quad (3.42)$$

The rigidity constants can then be rewritten to

$$D_1 = \Omega^2 D, \quad D_2 = 2\nu D, \quad D_3 = \Omega^{-2} D, \quad (3.43)$$

$$D_4 \approx \min \{ \Omega^2, \Omega^{-2} \} 2(1 - \nu) D.$$

The basic constants that are used to describe the material of isotropic and orthotropic plates are summarized in Tab. 3.1.

3.2.3. The impulse response of a rectangular plate

According to Troccaz et al. (2000), the impulse response of a rectangular plate at position (x, y) , subjected to a point force at excitation position (x_0, y_0) , is

$$w(x, y, t) = \sum_{n=0}^N \sum_{m=0}^M \underbrace{\frac{\Theta_{mn}(x, y) \Theta_{mn}(x_0, y_0)}{\omega_{r,mn} M_{mn}}}_{\text{amplitude } A_{mn}} \cdot \underbrace{e^{-\alpha_{mn} t}}_{\text{envelope}} \cdot \underbrace{\sin(\omega_{r,mn} t)}_{\text{sine wave}} \quad (3.44)$$

with the modal mass

$$M_{mn} = m_s \int_0^{l_y} \int_0^{l_x} \Theta_{mn}^2 dx dy, \quad (3.45)$$

where $m_s = \rho h$ is the plate mass per unit area (Putra and Thompson 2010). Θ_{mn} are the shapes of the individual modes which resonate at angular frequency $\omega_{r,mn}$. All damping is combined in the

3. Physical modeling for plausible auditory augmentation

decay factors α_{mn} . The indices m and n describe the mode indices in x - and y -direction, respectively.

Note that this impulse response describes the vibration of the plate itself and does not account for the radiated or perceived sound.

3.2.4. Mode shapes

The two-dimensional mode shapes $\theta_{mn}(x, y)$ are computed numerically from the one-dimensional characteristic beam functions $\theta_m(x)$ in x -direction and $\theta_n(y)$ in y -direction, respectively. The 2D mode shape is the Cartesian or outer product of the two characteristic beam functions:

$$\Theta_{mn}(x, y) = \theta_m^T(x) \theta_n(y) . \quad (3.46)$$

Approximate solutions for the characteristic beam functions are provided by Warburton (1954) and summarized in App. A.1 for all combinations of the three basic boundary conditions: hinged [h], clamped [c], and free [f]. These can be combined in 6 different ways: [hh], [cc], [ff], [cf], [ch], and [fh]. Note that, e.g., [hf] and [fh] are equivalent, and one can be obtained by spatial inversion of the other. While the characteristic beam functions for hinged and clamped boundary conditions are accurate, those with free edges are only approximate.

The mode shapes themselves are independent of the actual plate dimensions and can thus be pre-computed in normalized coordinates. What changes with dimensions, however, is their scaling and thus the corresponding natural frequencies.

3.2.5. Undamped natural frequencies

The mode shapes Θ from the previous section are now required for computing the plate's resonant frequencies. In combination with the elastic constants D_1 - D_4 and dimensions ($l_x \times l_y \times h$), they describe the maximum potential energy E_{pot} of the plate (McIntyre and Woodhouse 1988; Warburton 1954), as shown in Eq. 3.47.

$$E_{\text{pot}} = \frac{h^3}{2} \int_{y=0}^{l_y} \int_{x=0}^{l_x} \left[D_1 \left(\frac{\partial^2 \Theta}{\partial x^2} \right)^2 + D_2 \frac{\partial^2 \Theta}{\partial x^2} \frac{\partial^2 \Theta}{\partial y^2} + D_3 \left(\frac{\partial^2 \Theta}{\partial y^2} \right)^2 + D_4 \left(\frac{\partial^2 \Theta}{\partial x \partial y} \right)^2 \right] dx dy \quad (3.47)$$

On the other hand, the maximum kinetic energy E_{kin} is connected to the modal mass M_{mn} from

Eq. 3.45 through the angular frequency $\omega = 2\pi f$:

$$E_{\text{kin}} = \frac{\omega^2}{2} M_{mn} = \frac{\omega^2 \rho h}{2} \int_0^{l_y} \int_0^{l_x} \Theta_{mn}^2 dx dy . \quad (3.48)$$

After Rayleigh's principle, if the total mechanical energy $E_{\text{pot}} + E_{\text{kin}}$ stays constant, $E_{\text{pot}}/E_{\text{kin}} = 1$ applies for the undamped natural frequencies ω_0 , and thus

$$\omega_0^2 = 2E_{\text{pot}}/M_{mn} . \quad (3.49)$$

For the purpose of inserting Eq. 3.47 and 3.48 in Eq. 3.49 we define 4 coefficients:

$$\begin{aligned} G_x^4 &= \frac{\iint \left(\frac{\partial^2 \Theta}{\partial x^2} \right)^2 dx dy}{\Psi} \\ H_x H_y &= \frac{\iint \frac{\partial^2 \Theta}{\partial x^2} \frac{\partial^2 \Theta}{\partial y^2} dx dy}{\Psi} \\ G_y^4 &= \frac{\iint \left(\frac{\partial^2 \Theta}{\partial y^2} \right)^2 dx dy}{\Psi} \\ J_x J_y &= \frac{\iint \left(\frac{\partial^2 \Theta}{\partial x \partial y} \right)^2 dx dy}{\Psi} \end{aligned} \quad (3.50)$$

$$\text{with } \Psi = \pi^4 \iint \Theta^2 dx dy = \pi^4 \frac{M_{mn}}{m_s} . \quad (3.51)$$

Eq. 3.49 thus becomes

$$\omega_0^2 = \frac{\pi^4 h^2}{\rho S^2} \left[\frac{D_1}{r_a^2} G_x^4 + D_2 H_x H_y + D_3 r_a^2 G_y^4 + D_4 J_x J_y \right] \quad (3.52)$$

with surface area $S = l_x \cdot l_y$ and aspect ratio $r_a = l_x/l_y$. Note that $l_x \geq l_y$ and thus $r_a \geq 1$.

Analytic solutions to the factors G_x , H_x , and J_x , as well as G_y , H_y , and J_y are provided by Warburton (1954) for all combinations of the main boundary conditions, analog to the mode shapes in Sec. 3.2.4. Apart from that, they only depend on the number of nodal lines n and can thus be pre-computed. The equations are provided in App. A.2.

If Ω is separated from D_1 - D_4 , Eq.3.52 becomes

$$\begin{aligned} \omega_0 &= \pi^2 \frac{h}{S} \sqrt{\frac{D}{\rho}} \left[\frac{\Omega^2}{r_a^2} G_x^4 + 2\nu H_x H_y + \frac{r_a^2}{\Omega^2} G_y^4 + \min \left\{ \Omega^2, \frac{1}{\Omega^2} \right\} 2(1-\nu) J_x J_y \right]^{1/2} \end{aligned} \quad (3.53)$$

The 1st term in Eq. 3.53 scales the resonance frequencies altogether:

$$2\pi \cdot \Phi = 2\pi \cdot \frac{\pi h}{2S} \sqrt{\frac{D}{\rho}} = 2\pi \frac{\pi h}{\sqrt{48} S} c_L \quad (3.54)$$

with the longitudinal wave velocity

$$c_L = \sqrt{\frac{E}{\rho(1-\nu^2)}} = \sqrt{\frac{12D}{\rho}} \quad (3.55)$$

The 2nd term describes the non-dimensional frequency factors between the individual modes,

$$\lambda = \left[r_p^2 G_x^4 + 2\nu H_x H_y + \frac{1}{r_p^2} G_y^4 + \min\left\{\Omega^2, \frac{1}{\Omega^2}\right\} 2(1-\nu) J_x J_y \right]^{1/2} \quad (3.56)$$

with the spectral shaping factor $r_p = \Omega/r_a$. Orthotropy and aspect ratio thus have an oppositional effect on the frequency factors. r_p spreads the frequency factors symmetrically for values other than 1.

The lowest frequency factors are usually connected to the length of the plate and thus G_x^4 . In order to achieve a notation that describes frequency factors with respect to the usually lowest mode 1/0, r_p can be moved into Φ , leading to

$$2\pi \cdot \hat{\Phi} = 2\pi \frac{\pi h}{\sqrt{48} S} r_p c_L \quad (3.57)$$

and

$$\hat{\lambda} = \left[G_x^4 + \frac{1}{r_p^2} 2\nu H_x H_y + \frac{1}{r_p^4} G_y^4 + \frac{1}{r_p^2} \min\left\{\Omega^2, \frac{1}{\Omega^2}\right\} 2(1-\nu) J_x J_y \right]^{1/2} \quad (3.58)$$

The frequency in Hz then becomes

$$f_0 = \frac{\omega_0}{2\pi} = \frac{\Phi \lambda}{2\pi} = \frac{\hat{\Phi} \hat{\lambda}}{2\pi} \quad (3.59)$$

In the isotropic case, Eq. 3.56 and 3.58 simplify to

$$\hat{\lambda}|_{\Omega=1} = \left[\frac{1}{r_a^2} G_x^4 + 2\nu H_x H_y + r_a^2 G_y^4 + 2(1-\nu) J_x J_y \right]^{1/2} \quad \text{and} \quad (3.60)$$

$$\hat{\lambda}|_{\Omega=1} = \left[G_x^4 + r_a^2 2\nu H_x H_y + r_a^4 G_y^4 + r_a^2 2(1-\nu) J_x J_y \right]^{1/2} \quad (3.61)$$

and 3.57 simplifies to

$$\hat{\Phi}|_{\Omega=1} = \frac{\pi h}{\sqrt{48} S} \frac{1}{r_a} c_L = \frac{\pi h}{\sqrt{48} l_x^2} c_L \quad (3.62)$$

3.2.6. Damping

One of the most difficult aspects to compute is the plate's damping. And it is even impossible for some damping mechanisms. However, there are ways to roughly estimate frequency-dependent damping coefficients from physical parameters of the plate. One of the most precise damping models for rectangular plates has been presented by Chaigne and Lambourg (2001).

Under the condition of small damping, which means that the Q-factors of the individual modes are much greater than unity, Chaigne and Lambourg (after McIntyre and Woodhouse 1988) proposed to expand the four rigidity constants D_i by complex perturbation terms $\tilde{d}_i(j\omega)$, to form complex rigidities

$$\tilde{D}_i(j\omega) = D_i (1 + \tilde{d}_i(j\omega)) \quad (3.63)$$

with $i = \{1, 2, 3, 4\}$. The imaginary part of $\tilde{d}_i(j\omega)$ then provides a good approximation of the four small quantities η_i ,

$$\eta_i = \frac{\Im\{\tilde{D}_i(j\omega)\}}{\Re\{\tilde{D}_i(j\omega)\}} \approx \Im\left\{\frac{\tilde{D}_i(j\omega)}{D_i}\right\} \approx \Im\{\tilde{d}_i(j\omega)\} \quad (3.64)$$

which, via the weighting factors $J_{i,mn}$, sum up to the overall loss factors

$$\eta_{mn} = \sum_{i=1}^4 \eta_i J_{i,mn} \quad (3.65)$$

for the individual modes. Note that the frequency-dependency sustains but is omitted for readability. From here on, also the indices m and n will be omitted.

The weighting factors are nothing else than the normalized coefficients G , H , and J from Sec. 3.2.5:

$$J_i = \frac{J'_i}{\sum_{i=1}^4 J'_i} \quad (3.66)$$

3. Physical modeling for plausible auditory augmentation

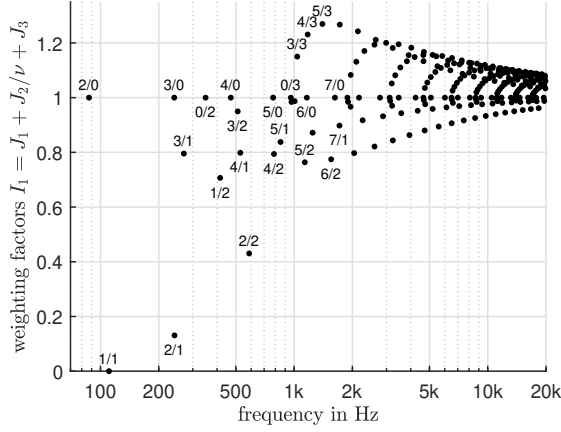


Figure 3.12.: Thermoelastic weighting factors I_1 of a rectangular metal plate with free boundaries, $\nu=0.3$, and $r_a=2$.

with

$$\begin{aligned} J'_1 &= \Omega^2 G_x^4, & J'_2 &= 2\nu H_x H_y, \\ J'_3 &= \frac{1}{\Omega^2} G_y^4, & J'_4 &= \frac{G_{xy}}{3D} J_x J_y. \end{aligned} \quad (3.67)$$

In the isotropic case, these simplify to

$$\begin{aligned} J'_1|_{\Omega=1} &= G_x^4, & J'_2|_{\Omega=1} &= 2\nu H_x H_y, \\ J'_3|_{\Omega=1} &= G_y^4, & J'_4|_{\Omega=1} &= 2(1-\nu) J_x J_y, \end{aligned} \quad (3.68)$$

and Eq. 3.65 reduces to

$$\eta_{mn}|_{\Omega=1} = \eta_1 I_{1,mn} + \eta_4 I_{4,mn} \quad (3.69)$$

with

$$\begin{aligned} I_{1,mn} &= J_{1,mn} + J_{3,mn} + \frac{1}{\nu} J_{2,mn} \quad \text{and} \\ I_{4,mn} &= J_{4,mn} - 2(1-\nu) J_{2,mn}. \end{aligned} \quad (3.70)$$

The J s simply indicate the partitioning of potential energy between the four rigidities, and therefore always sum up to 1 to ensure the conservation of energy (McIntyre and Woodhouse 1988):

$$\sum_{i=1}^4 J_{i,mn} = 1. \quad (3.71)$$

The values of I_1 for the individual modes of a free metal plate with $\nu=0.3$ and $r_a=2$ are shown in Fig. 3.12.

The above equations showed how the loss factors η are weighted by the individual mode shapes

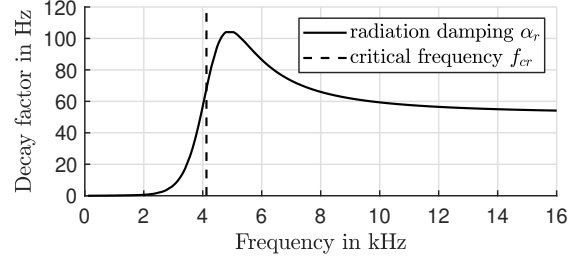


Figure 3.13.: Radiation damping α_r and critical frequency f_{cr} of an aluminum plate ($h = 3$ mm thick).

and material, but the actual amount of loss is still unknown. The computation of the loss factors needs to be carried out separately for each individual damping mechanism. These include external damping due to radiation or viscous mounting, and also intrinsic damping due to thermoelasticity or viscoelasticity. In the following sections, the corresponding loss factors are derived individually for each mechanism. The individual loss factors can be summed up directly to form the total loss of the plate, as shown in the end in Sec. 3.2.6.5.

3.2.6.1. Radiation damping

The largest source of damping is the radiation of sound itself. The sound we hear describes a transfer of energy from the physical object to our ears. The better a certain mode is radiated, the stronger it is damped. This tight connection with radiation efficiency is further described in Sec. 3.2.10; we will now first describe the radiation damping itself.

For isotropic plates, the damping due to radiation was formulated by Chaigne and Lambourg (2001) to be

$$\eta_r \approx \Im \{ \tilde{d}_r(j\omega) \}, \quad (3.72)$$

with

$$\tilde{d}_r(j\omega) = \frac{2}{\omega_{cr}} \frac{\rho_0 c_0}{\rho h} \frac{\sum_{m=1}^3 b_{r,m} \left(\frac{j\omega}{\omega_{cr}} \right)^m}{\sum_{n=0}^3 a_{r,n} \left(\frac{j\omega}{\omega_{cr}} \right)^n} \quad (3.73)$$

and the material constants from Tab. 3.2.

Above its cutoff frequency, the critical frequency

$$\omega_{cr} = \frac{c_0^2}{h} \sqrt{\frac{\rho}{D}} \quad \text{or} \quad f_{cr} = \frac{1}{2\pi} \frac{c_0^2}{h} \sqrt{\frac{\rho}{D}}, \quad (3.74)$$

radiation is the predominant source of loss (see Fig. 3.13). Below that frequency, other mechanisms such as thermoelastic, viscoelastic, and viscous damping are much larger.

Table 3.2.: Material constants for radiation damping.

constant	value	units	description
ρ_0	1.2	kg m^{-3}	density of the air
c_0	344	m s^{-1}	speed of sound in air
b_r	[0.0620 0.5950 1.0272]	–	–
a_r	[1.1669 1.6574 1.5528 1]	–	–

Table 3.3.: Material constants for thermoelastic damping.

constant	description	units
κ	thermal conductivity	$\text{W m}^{-1} \text{K}^{-1}$
C	specific heat at constant strain	$\text{J kg}^{-1} \text{K}^{-1}$
T_0	absolute temperature; usually 294.15 K (21 °C)	K
α_T	thermal expansion coefficient	K^{-1}

The loss factor η_r completely describes the radiation damping without the need to be scaled by J_{mn} or I_{mn} .

3.2.6.2. Thermoelastic damping

Any vibration of a rigid object implies some kind of temporary deformation or bending. Some parts are therefore compressed and thus become hotter, while others are extended and thus become cooler. As the object vibrates, the temperature at a vibration maximum will oscillate between hot and cold. If the material does not conduct any heat at all, then no energy is lost. If, however, the local heat is conducted to neighboring regions and thus spreads across the plate, the corresponding energy is lost forever and the plate heats up as a whole.

For such thermoelastic losses, Chaigne and Lam-bourg (2001) derived the complex perturbation $\tilde{d}_{it}(j\omega)$ to equal

$$\tilde{d}_{it}(j\omega) = \frac{j\omega R_{it}}{j\omega + 1/\tau_t} \quad (3.75)$$

with

$$\begin{aligned} R_{1t} &= \frac{8T_0\phi_x^2}{\pi^4 D_1 \rho C}, \quad R_{2t} = \frac{16T_0\phi_x\phi_y}{\pi^4 D_2 \rho C}, \\ R_{3t} &= \frac{8T_0\phi_y^2}{\pi^4 D_3 \rho C}, \quad R_{4t} = 0, \end{aligned} \quad (3.76)$$

and, for the isotropic case,

$$\phi_x = \phi_y = \phi = \alpha_T \frac{E}{1 - 2\nu} \quad (3.77)$$

with thermal expansion coefficient α_T , and thermoelastic decay factor τ_t ,

$$\tau_t = \frac{\rho C h^2}{\kappa \pi^2} = \frac{h^2}{c_{1t}}, \quad (3.78)$$

with

$$c_{1t} = \frac{\kappa \pi^2}{\rho C}. \quad (3.79)$$

C is the specific heat at constant strain, κ is the thermal conductivity, and T_0 is the absolute temperature (see Tab. 3.3 for a summary of thermoelastic constants).

In practice, thermoelastic damping is relevant only in metals which are usually isotropic. In addition, with $R_{4t} = 0$, there is no thermoelastic loss for modes that are only connected to D_4 and hence to the shear modulus G_{xy} . Thermoelastic damping thus only affects $\tilde{d}_1(j\omega)$:

$$\tilde{d}_{it}(j\omega) \approx \tilde{d}_{1t}(j\omega) = \frac{j\omega R_{1t}}{j\omega + 1/\tau_t} \quad (3.80)$$

with

$$R_{1t} = \frac{8T_0\phi^2}{\pi^4 D \rho C}. \quad (3.81)$$

The thermoelastic loss factor η_{1t} therefore becomes

$$\begin{aligned} \eta_{1t} &\approx \Im \{ \tilde{d}_{1t}(j\omega) \} = \frac{R_{1t}}{\tau_t \omega + 1/\tau_t \omega} \\ &= \frac{R_{1t}}{\frac{h^2 \omega}{c_{1t}} + \frac{c_{1t}}{h^2 \omega}}. \end{aligned} \quad (3.82)$$

For aluminum, the thermoelastic damping coefficients are $R_{1t} = 8.45 \times 10^{-3} \text{ rad m}^2 \text{ s}^{-1}$ and

3. Physical modeling for plausible auditory augmentation

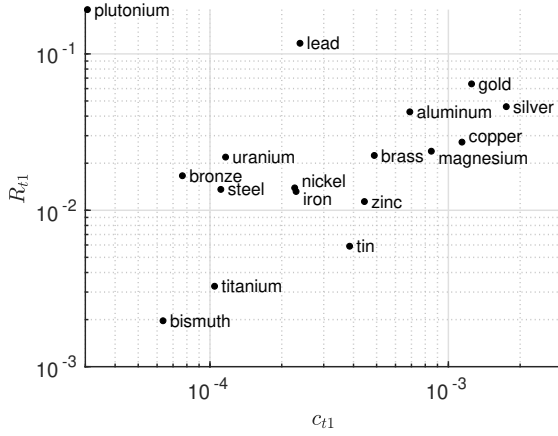


Figure 3.14.: Relationship between the thermoelastic damping coefficients R_{1t} and c_{1t} for different metals. The values are derived from the basic material constants.

$c_{1t} = 8.0 \times 10^{-4} \text{ rad m}^2 \text{ s}^{-1}$ (Chaigne and Lambourg 2001). As Fig. 3.14 shows, there is no simple correlation between these two coefficients for different metals.

Figure 3.15 shows that instead of a constant loss factor similar to viscoelastic damping, we obtain an almost constant decay factor α_t for thermoelastic damping. While viscoelastic losses are assumed to be independent of plate dimensions, the thermoelastic losses depend on h^2 . For realistic frequencies above 50 Hz and thicknesses above 3 mm, the frequency-dependency of α_t can be considered negligible. The total thermoelastic loss finally equals

$$\eta_t = \eta_{1t} I_1 \quad (3.83)$$

The thermoelastic damping of an exemplary aluminum plate is depicted in Fig. 3.16, together with the radiation damping and the viscoelastic damping which will be explained in the following section.

3.2.6.3. Viscoelastic damping

Viscoelastic loss results from the combination of viscous and elastic properties of a material. If a purely elastic material is subject to deformation, it returns back to its original shape in exactly the same way. The viscosity adds hysteresis to the stress-strain curve due to its dependency on time. The area within this hysteresis loop equals the amount of energy that is lost in the form of heat. At the time of writing there exists no model that quantifies the

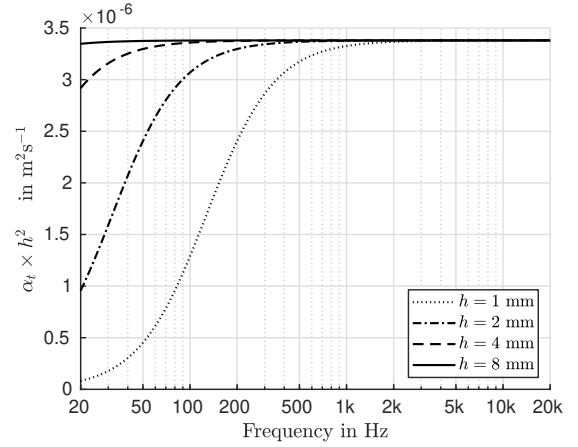


Figure 3.15.: Relationship between thermoelastic damping, thickness, and frequency in case of an aluminum plate.

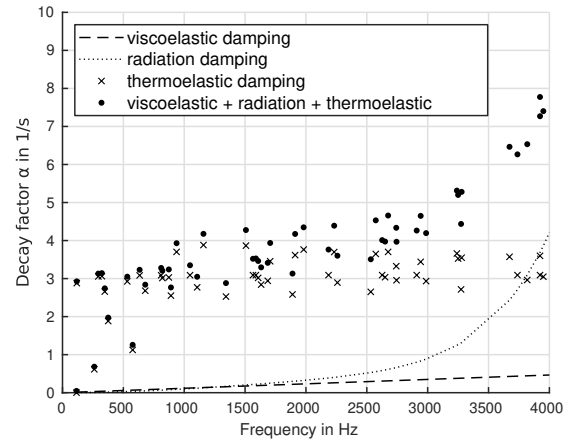


Figure 3.16.: The decay factors of an aluminum plate due to the three main damping mechanisms: radiation damping, thermoelastic damping, and viscoelastic damping.

viscoelastic losses of a plate without just measuring it (Zoghaib and Mattei 2015).

At least the rough characteristics of the viscoelastic damping is provided by Chaigne and Lambourg (2001), based on a so-called generalized Maxwell (or Wiechert–Maxwell) model (Gutierrez-Lemini 2014, pp. 81–83):

$$\tilde{d}_{iv}(j\omega) = \sum_{k=1}^K \frac{j\omega R_{kiv}}{j\omega + s_{kv}} \quad (3.84)$$

According to Chaigne and Lambourg, a resolution

Table 3.4.: Material constants for viscoelastic damping.

Material	$R_v / 10^{-3}$	$s_v / 10^3 \text{ rad s}^{-1}$
glass	$\begin{bmatrix} 1.625 & - & - & 1.625 \\ 1.962 & - & - & 1.962 \end{bmatrix}$	$\begin{bmatrix} 5.18 & - & - & 5.18 \\ 55.10 & - & - & 55.10 \end{bmatrix}$
carbon	$\begin{bmatrix} 1.32 & 0 & 8.8 & 10.4 \\ 5.0 & 0 & 44.0 & 14.4 \end{bmatrix}$	$\begin{bmatrix} 10.1 & 0 & 2.5 & 2.27 \\ 94.0 & 0 & 0.07 & 40.0 \end{bmatrix}$
wood (spruce)	$\begin{bmatrix} 8.18 & 0 & 16.7 & 15.2 \\ 10.0 & 0 & 70.0 & 35.0 \end{bmatrix}$	$\begin{bmatrix} 3.2 & 0 & 1.1 & 1.75 \\ 50.2 & 0 & 0.052 & 50.2 \end{bmatrix}$

of $K=2$ is sufficiently accurate:

$$\tilde{d}_{iv}(j\omega) = \frac{j\omega R_{1iv}}{j\omega + s_{1v}} + \frac{j\omega R_{2iv}}{j\omega + s_{2v}}. \quad (3.85)$$

The viscoelastic loss factors thus become

$$\eta_{iv} \approx \Im \{ \tilde{d}_{iv}(j\omega) \}, \quad (3.86)$$

and because η_{2v} is always zero,

$$\eta_v = \eta_{1v}J_1 + \eta_{3v}J_3 + \eta_{4v}J_4 \quad (3.87)$$

for orthotropic, and

$$\eta_v = \eta_{1v}I_1 + \eta_{4v}I_4 \quad (3.88)$$

for isotropic materials.

Chaigne and Lambourg (2001) fitted coefficients on measurements with plates made of glass, carbon fibers, and wood (see Tab. 3.4).

For isotropic materials, especially if the coefficients corresponding to D_1 and D_4 are identical (e.g., glass in Tab. 3.4), the loss factor is almost constant, which leads to a direct proportionality between decay factor α_v and frequency. Average loss factors of common materials have been collected by Cremer et al. (2005, pp. 191, 195–196). Assuming that there is only one mechanism of loss, as is approximately the case for viscoelastic loss below the critical frequency, these may serve as a rough estimate. Figure 3.17 shows the viscoelastic decay factor for glass, as measured by Chaigne and Lambourg, in comparison with the range $\eta_v = [0.6, 2] \times 10^{-3}$ that is given by Cremer et al. (2005, p. 195). The geometric mean of this range is already an incredibly good fit to the measurements by Chaigne and Lambourg.

Also Zoghaib and Mattei (2015) use a constant loss factor to model the presumably viscoelastic losses of an aluminum plate, that their predicted thermoelastic and radiation losses could not account for. Their value of $\eta_v = 3.7 \times 10^{-4}$ resides a bit above

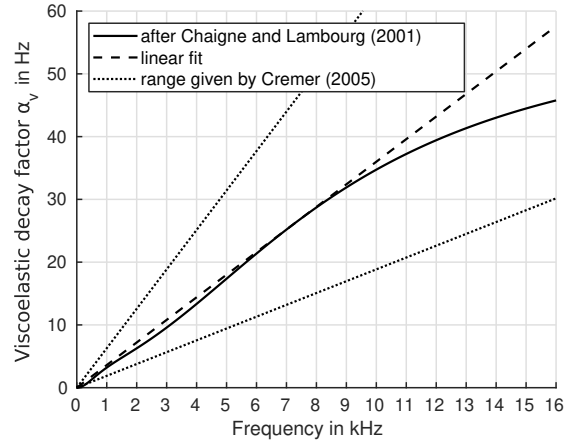


Figure 3.17.: Viscoelastic decay factor α_v for glass, as a function of frequency.

the range of $\eta_v = [0.03, 1] \times 10^{-4}$ that Cremer provides for aluminum. The contribution of viscoelastic damping in case of an aluminum plate is visualized in Fig. 3.16. The fact that thermoelastic damping is almost zero for some modes basically suppresses its effect on a statistically measured loss factor, which means that the values from Cremer might actually only capture viscoelastic losses. Despite the divergence between different models, average loss factors such as from Cremer can still be considered as a solid estimate of viscoelastic loss below the critical frequency.

For wood, Ono and Norimoto (1983) analyzed a huge dataset of measurements of internal friction, and found a “standard trend”, based on only Young’s modulus E and density ρ (see also Brémaud 2012b; Brémaud et al. 2013):

$$\eta_v \approx 10^{-1.23} \cdot \left(\frac{E}{\rho} \right)^{0.68} \quad (3.89)$$

More recent measurements by Brémaud (2012a) and Brémaud et al. (2013) suggest that the loss

3. Physical modeling for plausible auditory augmentation

factor of so-called “compression wood” is slightly lower, while the loss factor of normal, fast-grown wood is slightly higher than the standard trend. This difference is attributed to the crystalline microfibrils of cellulose which interconnect the main fibers. The larger the microfibril angle, i.e., the angle between microfibrils and grain direction, the larger the loss factor. Compression wood usually appears only in one or a few rings, mostly in branches and leaning stems, in reaction to deformation due to gravity. In the search for a simple model for wood as such, the standard trend, however, seems to be a good compromise.

Figure 3.18 illustrates the measured viscoelastic loss factors η_v of several materials as a function of the longitudinal wave velocity c_L . On the logarithmic scale, it seems that there is not much difference between the standard trend and a simple indirect proportionality with c_L . We infer a rough proportionality factor of 57 for wood and plastic, 5.7 for glass, and 0.57 for metal. This simple proportionality is preferred over a precise fit, as it implies that a traveling wave receives a constant amount of damping per distance traveled, independent of its velocity.

Among the few viscoelastic loss factors we can actually trust are those of glass (Chaigne and Lambourg 2001) and aluminum (Zoghaib and Mattei 2015), as these are based on laser vibrometer measurements and could be freed from other damping mechanisms. For anisotropic materials, not even laser vibrometer measurements allowed a separation of damping mechanisms (Chaigne and Lambourg 2001). As furthermore acrylic and wood sounded somehow overly damped in informal listening tests, we generalize the model even further to $\eta_{v,M} = 5.7/c_L$ for non-metals and $\eta_{v,M} = 0.57/c_L$ for metals.

3.2.6.4. Viscous damping

As in all the above models for thermoelastic, viscoelastic, and radiation damping, the decay factor approaches zero at low frequencies, Chaigne and Lambourg (2001) introduced a constant viscous decay factor α_f . While the actual values were simply fitted to the measurements, without further investigations, they actually account for the suspension of the plate. By the restraining filaments, the translational and rotational movements, i.e., the 0 Hz modes 0/0, 1/0, and 0/1, are suppressed. In general, viscous damping not only includes the loss due to external damping devices such as ab-

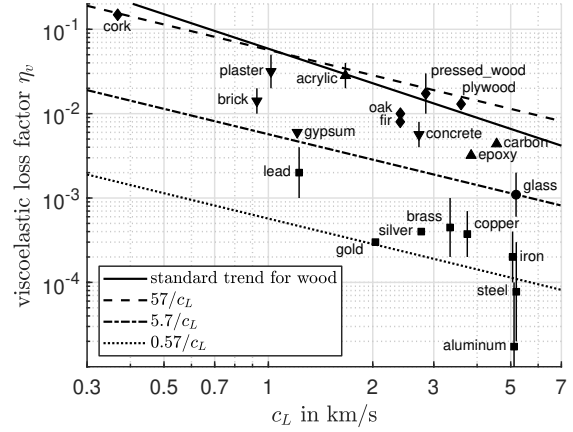


Figure 3.18.: Viscoelastic loss factor η_v as a function of longitudinal wave velocity c_L . Vertical lines represent the range between minimum and maximum value from the literature, while markers represent their geometric mean and gross material category: \blacklozenge =wood, \blacksquare =metal, \blacktriangle =plastic, \blacktriangledown =stone, \bullet =glass.

sorbers, but also the loss due to the viscosity of the surrounding material. Values of α_f are given by Chaigne and Lambourg (2001) for aluminum (0.032 Hz), glass (0.88 Hz), carbon (0.8 Hz), and spruce wood (2.4 Hz).

We use a constant viscous damping α_f as a simple way to model the losses that are attributed to the mounting of the plate or any other damping material that is attached to it, including interaction such as damping with the palm of the hand. Decay times in the low-frequency range are thus limited, which is common practice in perceptual studies (e.g., McAdams et al. 2004; McAdams et al. 2010).

3.2.6.5. Complete damping model

After McAdams et al. (2010), materials between aluminum and glass can be effectively simulated by blending between thermoelastic and viscoelastic damping with the help of a damping interpolation parameter H . We generalize this approach and interpret H as the metallicity of a material. The viscoelastic loss factor η_v therefore becomes

$$\eta_v = (1 - H)\eta_{v,M} + H\eta_{v,M} , \quad (3.90)$$

and the overall damping of the plate sums up to

$$\eta = (1 - H)\eta_{v,M} + H(\eta_{v,M} + \eta_t) + \eta_r + \frac{\alpha_f}{2\omega} . \quad (3.91)$$

The overall decay factors of an exemplary aluminum plate are shown in Fig. 3.19. The complete damping comprises all the previously discussed

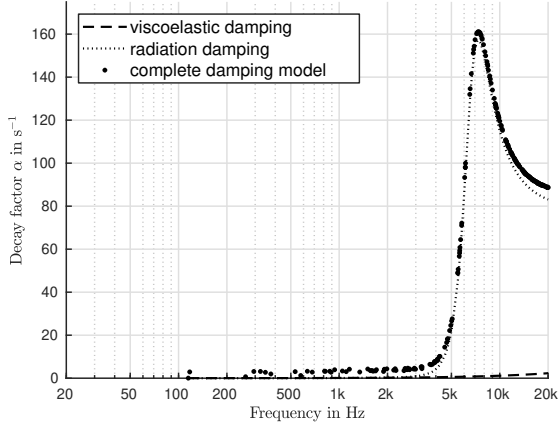


Figure 3.19.: The overall decay factors of an aluminum plate, in comparison with the individual contributions from radiation damping and viscoelastic damping.

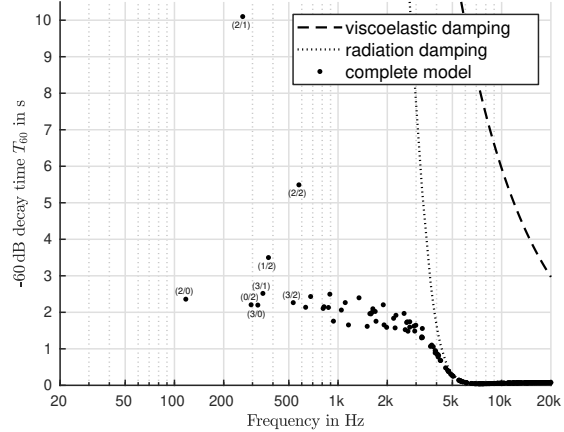


Figure 3.20.: The overall -60 dB decay times of an aluminum plate, in comparison with the individual contributions from radiation damping and viscoelastic damping.

damping mechanisms (radiation, viscoelastic, thermoelastic, and viscous damping). For comparison, Fig. 3.20 shows the resulting -60 dB decay times.

3.2.7. Modal weights due to the position of contact

Any object sounds different, depending on where the excitation occurs. This behavior is not random but rather a direct result of the mode shapes at the area of contact. Assuming an infinitesimal area of contact, the force that is injected into a single mode, and therefore its amplitude, is directly proportional to the value of the mode shape at the excitation position divided by the modal mass M_{mn} .

If the mode shapes are pre-computed for normalized dimensions, i.e., $l_x = l_y = 1$, the amplitude $A_{mn}(x_0, y_0)$ of mode m/n at the normalized excitation position (x_0, y_0) is returned via bilinear interpolation. While 4-point polynomial interpolation would give significantly more accurate results, the marginal perceptual difference is usually not worth the computational effort.

The amplitude weights for a plate with free boundary conditions, hit at the edge of the long side in the maximum of the $3/0$ mode, is shown in Fig. 3.21.

3.2.8. Hertz' law of contact

In the previous section we assumed a point-shaped excitation. Even under the assumption that there is no indentation of the mallet into the plate, the elastic deformation of plate and mallet lead to an

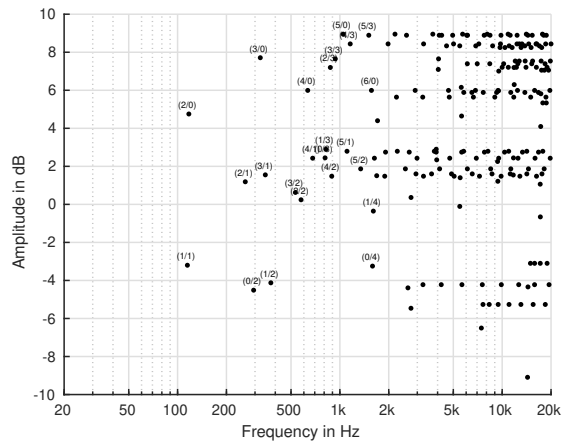


Figure 3.21.: The amplitude weights resulting from excitation of a free rectangular plate at the edge of the long side in the maximum of the $3/0$ mode.

area of contact that is larger than zero. In our collision model, we assume a perfectly flat surface that is impacted by a sphere.

As Schedin et al. (1999) proposed, this excitation force can be sufficiently approximated by a Hann window (also called raised cosine window):

$$F_{\text{ex}}(t) = \begin{cases} F_{\text{max}} \sin^2\left(\frac{\pi t}{T_{\text{max}}}\right), & \text{if } 0 < t < T_{\text{max}} \\ 0, & \text{otherwise} \end{cases} \quad (3.92)$$

with force maximum F_{max} and duration T_{max} . These sound parameters are connected to physical parameters via Hertz' law of contact. Chaigne and Doutaut

3. Physical modeling for plausible auditory augmentation

(1997) provide a detailed summary together with measurements on xylophone bars. The impact duration

$$T_{\max} = 3.2181 \left(\frac{\mu_{\text{BM}}^2}{K_{\text{M}}^2 V_0} \right)^{1/5} \quad (3.93)$$

is connected to the instantaneous velocity V_0 of the mallet, the combined stiffness coefficient K_{M} , and the reduced mass μ_{BM} of bar and mallet

$$\mu_{\text{BM}} = \frac{m_{\text{B}} m_{\text{M}}}{m_{\text{B}} + m_{\text{M}}} \quad (3.94)$$

where m_{M} is the effective mass of the mallet and m_{B} is mass of the bar. The combined stiffness depends on the mallet radius R_{M} and the combined rigidity D_{BM} according to

$$K_{\text{BM}} = \sqrt{R_{\text{M}}/D_{\text{BM}}} \quad (3.95)$$

and

$$D_{\text{BM}} = \frac{3}{4} \left(\frac{1 - \nu_{\text{B}}^2}{E_{\text{B}}} + \frac{1 - \nu_{\text{M}}^2}{E_{\text{M}}} \right) \quad (3.96)$$

with Young's moduli E_{B} and E_{M} , as well as Poisson's ratios ν_{B} and ν_{M} of bar and mallet, respectively. The maximum force occurs at the time of maximum compression and equals

$$F_{\max} = K_{\text{BM}} \delta_{\max}^{3/2} \quad (3.97)$$

with the maximum value of the compression

$$\delta_{\max} = \left(\frac{5}{4} \frac{\mu_{\text{BM}}}{K_{\text{BM}}} \frac{2^{5/2}}{V_0^{4/5}} \right) \quad (3.98)$$

If F_{\max} is given instead of V_0 , Eq. 3.97–3.98 can be reformulated to obtain

$$V_0 = K_{\text{BM}}^{-5/6} \left(\frac{5}{4} \frac{\mu_{\text{BM}}}{K_{\text{BM}}} \right)^{-1/2} \quad (3.99)$$

The resulting temporal Hann window can be applied to the excitation signal via convolution. For auditory augmentation, however, we assume that this kind of low-pass filtering is anyway included already in the raw excitation signal.

The maximum radius of the contact area can be expressed as

$$R_{\max} = \sqrt{\delta_{\max} R_{\text{M}}} \quad (3.100)$$

and is utilized to model a spatial Hann window for the computation of individual amplitude weights of the modes due to excitation (see Sec. 3.2.7).

Instead of directly evaluating the mode shapes at a distinct position to model an ideal point excitation, each mode shape is then weighted by the spatial Hann window Θ_{ex} . The integral over the surface area gives the amplitude weight

$$A_{mn} = \int_0^{l_y} \int_0^{l_x} F_{\max} \frac{\Theta_{\text{ex}}(x - x_0, y - y_0)}{\iint \Theta_{\text{ex}} dx dy} \odot \Theta_{mn}(x, y) dx dy \quad (3.101)$$

where \odot denotes an element-wise or Hadamard product. The spatial Hann window follows the equation

$$\Theta_{\text{ex}}(x, y) = \begin{cases} \cos^2\left(\pi \frac{r}{R_{\max}}\right), & \text{if } 0 < r < R_{\max} \\ 0, & \text{otherwise} \end{cases} \quad (3.102)$$

with $r = \sqrt{x^2 + y^2}$.

3.2.9. Indentation hardness

Solid objects are usually not totally solid: depending on the actual material, the mallet caves into the object, leaving an indentation behind. The consequence is a larger area of contact which exceeds the deformation that is described by Hertz' law of contact. While it has no significant effect on modes with low spatial frequency (peaks larger than the contact area), modes with high spatial frequency (small peaks with respect to the contact area) are then barely excited. Speaking of sound, this leads to a low-pass filtering of the excitation signal. Similarly, the contact is stretched over time which again leads to a loss of high frequencies.

Hardness is actually not a material constant, but a property which is estimated via material probing. Depending on the used convention, some type of indenter is pressed into a sample of the material to create an indentation. If using a standardized pressure or force, then the size of the indentation is measured. Otherwise a standardized size of indentation is produced, and the needed amount of load is measured. In viscous materials such as plastics, also the duration of indentation plays a role and needs to be defined. In common hardness tests (e.g., Brinell hardness, Rockwell hardness, and Vickers hardness) the size of the indentation at a standardized and constant force is measured. In the wood sector, however, the most widespread procedure is Janka hardness testing, which measures the force that is

required to embed a metal ball halfway into the sample.

We use Brinell hardness, because it uses a spherical indenter and conceptually shares some similarity with our excitation model based on Hertz' law of contact. The Brinell Hardness HB in kilogram-force² (kgf) is defined as

$$HB = \frac{P}{2\pi R_1(R_1 - \sqrt{R_1^2 - r_1^2})}, \quad (3.103)$$

with the radius of the indentation r_1 in mm, the radius of the indenter R_1 in mm, and the applied load P in kgf. If HB is given in MPa, a division by the standard gravity $g = 9.80665 \text{ m/s}^2$ yields the value in kgf. In case of orthotropic materials, we always refer to the side hardness, i.e., the resistance to a load that is perpendicular to the plate surface.

For wood, the Janka hardness F_{Janka} in N is converted to Brinell scale via Eq. 3.103. With the usual ball radius of $R_1 = 5.64 \text{ mm}$, and $r_1 = R_1$, we get

$$HB \approx \frac{F_{\text{Janka}}}{2\pi g R_1^2} = 0.00051 F_{\text{Janka}}. \quad (3.104)$$

While this conversion looks mathematically correct at first glance, it is actually quite crude and the results of a Brinell hardness test might differ significantly. However, this rough approximation is just sufficient for the purpose of sound generation. From Sec. 3.2.8 we know already that the impact between sphere and plate can be approximated by a force in the shape of a Hann- or raised-cosine window. In order to avoid a convolution with the Hann window itself, the frequency response is approximated by a 3rd-order low-pass filter, with -3 dB cutoff frequency f_{cH} (see Fig. 3.22):

$$f_{cH} \approx \frac{2.4733 \sqrt{\log(2)}}{\pi t_{\text{Hann}}} \quad (3.105)$$

The window duration and thus cutoff frequency depends on the physical parameters of sphere and plate, as well as their relative velocity.

Chaigne and Doutaut (1997) measured the impact force over time between different mallets and xylophone bars. Assuming that boxwood is much harder than a rosewood bar, the interaction duration of a boxwood mallet and a rosewood bar gives an approximate duration of $156 \mu\text{s}$ for rosewood. Further assuming that a rubber sphere applies the same

²kilogram-force: a deprecated gravitational metric unit of force; $1 \text{ kgf} = 9.80665 \text{ N}$

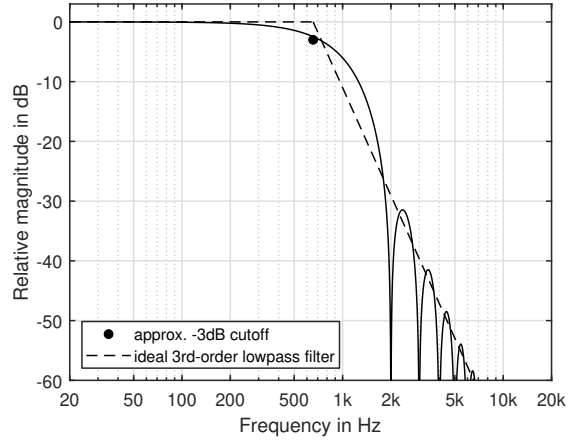


Figure 3.22.: The frequency response of a Hann window, and its approximation by an ideal 3rd-order low-pass filter.

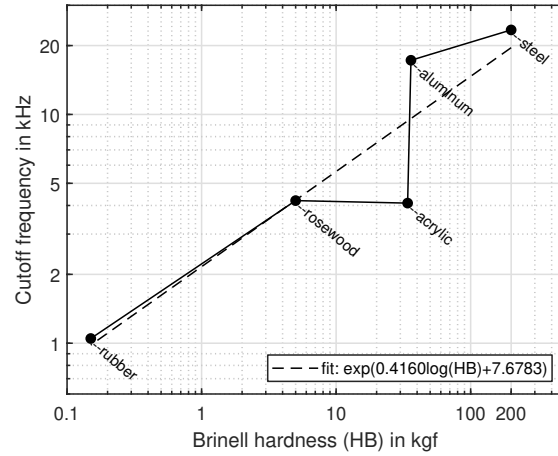


Figure 3.23.: Upper cutoff frequency f_{cH} vs. Brinell hardness HB .

force to an aluminum plate as would an aluminum sphere to a rubber plate, the interaction duration between rubber mallet and aluminum bar gives an estimated window duration of $625 \mu\text{s}$ for rubber.

Troccaz et al. (2000) made similar measurements with a steel sphere and plates made of acrylic, aluminum, and steel, leading to window lengths of about $160 \mu\text{s}$, $38 \mu\text{s}$, and $28 \mu\text{s}$, respectively.

Inserting these window durations into Eq. 3.105 yields a cutoff frequency for each of the corresponding materials. In Fig. 3.23, these cutoff frequencies are plotted against the Brinell hardness numbers, respectively. A simple linear regression (on a logarithmic scale) seems to be a good approximation for the effect of hardness on the upper cutoff frequency:

$$\log(f_{cH}) \approx 0.4160 \log(HB) + 7.6783 \quad (3.106)$$

3.2.10. Radiation efficiency

A rectangular plate doesn't radiate its sound equally across all frequencies and to all directions. The overall radiation can be regarded as the combined radiation of the individual modes. Each mode creates oscillating maxima and minima of sound pressure, according to its individual shape. The surrounding air creates an acoustic shortcut between pressure maxima and their neighboring pressure minima. An additional acoustic shortcut occurs for an un baffled plate through air flowing around the edge between both sides of the plate. In both cases, the vibration is not radiated as sound pressure in normal direction of the plate but rather leads to oscillating airflow that is parallel to the plate or circular around it. This behavior is different for each mode and thus frequency. Together with constructive and destructive interference between radiation from different locations on the plate, a frequency-, position- and direction-dependent radiation pattern is formed. For the purpose of auditory augmentation, the simulation of a realistic radiation pattern is not necessary. However, the frequency-dependent radiation efficiency is essential for the typical sound of the plate.

An empirical model for the radiation efficiency of an un baffled plate has been derived by Putra and Thompson (2010). In general, the radiation efficiency σ is defined as

$$\sigma = \frac{P_{\text{rad}}}{\rho_0 c_0 S \langle v^2 \rangle} \quad (3.107)$$

with the radiated sound power P_{rad} , air density ρ_0 , sound velocity in air c_0 , the surface area of the plate S , and the spatially averaged mean-square normal velocity of the plate $\langle v^2 \rangle$.

Putra and Thompson obtained an approximation by a set of empirical formulae. They define a few edge frequencies which depend on physical parameters of the plate, and divide the radiation efficiency into segments:

$$f_e = \frac{c_0}{2\sqrt{S}} \quad (3.108)$$

$$f_{cr} = \frac{1}{2\pi} \frac{c_0^2}{h} \sqrt{\frac{\rho}{D}} \quad (3.109)$$

$$f_b = \sqrt{f_1 f_2} \quad (3.110)$$

where f_1 and f_2 are the lowest two natural frequencies of the plate. In case of simply-supported edges, $f_1 = f_{1,1}$ and $f_2 = f_{2,1}$. The segments of the radiation efficiency are

$$\sigma_1 = \frac{4S^2 f^4}{c_0^4} \quad \text{for } f < f_b \quad (3.111)$$

$$\sigma_2 = \sigma_e \left(\frac{f}{f_e} \right)^2 \quad \text{for } f_2 < f \leq f_e \quad (3.112)$$

$$\sigma_e = \frac{p c_0}{4\pi^2 S f_{cr}} \cdot \frac{\psi}{(1 - \psi^2)^2} + \left(\pi \eta \frac{f}{f_{cr}} \right)^{3/2} \quad (3.113)$$

for $f_e < f < f_{cr}$, with perimeter $p = 2(l_x + l_y)$ and $\psi = \sqrt{f/f_{cr}}$, and

$$\sigma_c = \left(1 - \frac{f_{cr}}{f} \right)^{-1/2} \quad \text{for } f_e < f < f_{cr} \quad (3.114)$$

Between f_b and f_2 , the radiation efficiency is logarithmically interpolated:

$$\sigma_{12} = \sigma_2^\epsilon \sigma_1^{1-\epsilon} \quad (3.115)$$

with

$$\epsilon = \frac{f - f_b}{f_2 - f_b} \quad (3.116)$$

In summary, the "unlimited" radiation efficiency σ' is then:

$$\sigma' = \begin{cases} \sigma_1 & \text{for } f < f_b \\ \sigma_{12} & \text{for } f_b \leq f \leq f_2 \\ \sigma_2 & \text{for } f_2 < f \leq f_e \\ \sigma_e & \text{for } f_e < f < f_{cr} \\ \sigma_c & \text{for } f \geq f_{cr} \end{cases} \quad (3.117)$$

The final radiation efficiency σ is limited to become

$$\sigma = \min \{ \sigma', \sigma_{\text{max}} \} \quad (3.118)$$

with

$$\sigma_{\text{max}} = \left(0.5 - \frac{0.15}{r_a} \right) \sqrt{\frac{2\pi f_{cr} l_y}{c_0}} \quad (3.119)$$

Note that Eq. 3.119 is only valid for aspect ratios between 1 and 5. The final radiation efficiency of an exemplary aluminum plate is shown in Fig. 3.24. There is an obvious connection between radiation damping and radiation efficiency. Figure 3.25 illustrates that $\sigma \approx (\rho h / \rho_0 c_0) \alpha_r$, with $\rho_0 c_0$ being the characteristic specific acoustic impedance.

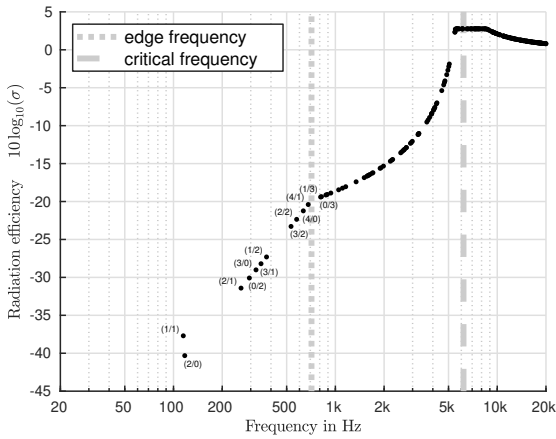


Figure 3.24.: The radiation efficiency of an aluminum plate.

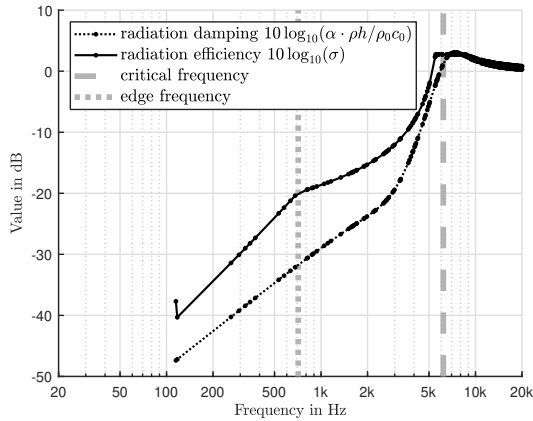


Figure 3.25.: Relationship between radiation damping and radiation efficiency of an aluminum plate.

3.3. Modal analysis and synthesis

In the previous sections, we learned how physical parameters of simple physical objects map to sound parameters. This section now briefly describes how these sound parameters are used for sound synthesis.

3.3.1. Modal analysis

Modal analysis is based on the assumption that the physical system is linear and time-invariant (LTI). Mathematically speaking, a linear system is (a) additive: $f(x + y) = f(x) + f(y)$, and (b) homogeneous: $f(ax) = a f(x)$. Given these assumptions and that the system doesn't change with time, each mode

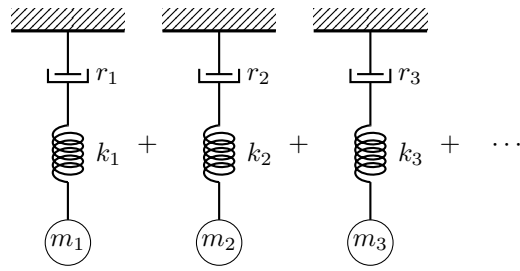


Figure 3.26.: Equivalent parallel model of modal masses m_n , modal stiffnesses k_n , and modal dampings r_n .

can be equivalently represented by a single harmonic oscillator with modal mass m_n , modal stiffness k_n , and modal damping r_n (see Fig. 3.26). These N simple harmonic oscillators which vibrate in parallel and completely independent from each other, form in sum an equivalent network for the original system of N coupled oscillators. Modal analysis basically neglects the original system network. Instead, all aspects of the system are only defined through its modes. These can be measured from the observed physical object itself, or computed from its physical properties, as done in the previous sections.

Regarding measurement, a rough estimate of the resonant frequencies and amplitudes is obtained from peak-picking the local maxima of the (smoothed) frequency response. The frequencies can be fine-tuned by method of adjustment where the frequency of a driving oscillator is manually tuned to the value which produces a maximum amplitude. If the driving oscillator is abruptly stopped, the time constant of the damping can be directly inferred from the envelope of the recorded decay.

3.3.2. Modal synthesis

Modal synthesis is the approach of modeling the sound of a rigid physical object through its modes. Instead of modeling the original system of coupled oscillators, only the equivalent model of parallel oscillators is recreated. Its impulse response $h(t)$ is defined by the sum of N modes, each modeled by an exponentially decaying sinusoid via three sound parameters: frequency f_n , amplitude A_n , decay factor α_n , and phase ϕ_n (see Eq. 3.120).

$$h(t) = \sum_{n=1}^N \underbrace{A_n}_{\text{amplitude}} \cdot \underbrace{e^{-\alpha_n t}}_{\text{envelope}} \cdot \underbrace{\sin(2\pi f_n t + \phi_n)}_{\text{sine wave}} \quad (3.120)$$

3. Physical modeling for plausible auditory augmentation

Any input signal (driving force) can be convolved with this impulse response to obtain the output signal. The impulse response basically describes a filterbank of N parallel resonant filters. Due to the exponential decay, each filter has an infinite impulse response (IIR) which can be modeled by an IIR digital filter.

The driven harmonic oscillator of Sec. 3.1.3 can be interpreted as a linear filter with generic non-dimensional input and output signals $x(t)$ and $y(t)$, respectively. To ensure that input and output are scaled equally, the driving force is scaled accordingly so that $F_d(t) = kx(t)$. The differential equation Eq. 3.11 thus becomes

$$m\ddot{y}(t) + r\dot{y}(t) + ky(t) = kx(t) . \quad (3.121)$$

3.3.2.1. Finite difference approximation

The above differential equation (Eq. 3.121) can be discretized in time ($t \rightarrow n$) and rewritten to a difference equation with the help of approximate solutions for the 1st derivative by simple backward

$$\dot{y}[n] \approx \frac{y[n] - y[n-1]}{T} \quad (3.122)$$

or forward difference approximation

$$\dot{y}[n] \approx \frac{y[n+1] - y[n]}{T} , \quad (3.123)$$

or their average being central difference approximation

$$\dot{y}[n] \approx \frac{y[n+1] - y[n-1]}{2T} , \quad (3.124)$$

with the sampling period $T = 1/F_s$ at sampling rate F_s . The second derivative can be written as successive forward and backward differentiation, resulting in

$$\ddot{y}[n] \approx \frac{y[n+1] - 2y[n] + y[n-1]}{T^2} . \quad (3.125)$$

Choosing the central 1st order approximation, the difference equation is formulated as

$$\begin{aligned} & m \frac{y[n+1] - 2y[n] + y[n-1]}{T^2} \\ & + r \frac{y[n+1] - y[n-1]}{2T} \\ & + ky[n] = kx[n] . \end{aligned} \quad (3.126)$$

As physical systems are always causal, meaning that they do not depend on the future, the whole equation must be delayed by one sample. A reformulation

yields

$$\begin{aligned} y[n] = & \underbrace{\left(\frac{2T^2k}{Tr + 2m} \right)}_{b_1} x[n-1] \\ & - \underbrace{\left(2 \frac{T^2k - 2m}{Tr + 2m} \right)}_{a_1} y[n-1] \\ & - \underbrace{\left(- \frac{Tr - 2m}{Tr + 2m} \right)}_{a_2} y[n-2] \end{aligned} \quad (3.127)$$

which is a 2nd-order digital filter, written in 'direct form 1', with coefficients b_1 , a_1 , and a_2 and the generic transfer function

$$H(z) = g \frac{b_0 + b_1z^{-1} + b_2z^{-2}}{1 + a_1z^{-1} + a_2z^{-2}} \quad (3.128)$$

where gain $g=1$ and $b_0=b_2=0$. This approximation is equal to the physical system, if the sampling rate approaches infinity. For a practicable audio sampling rate of $F_s = 48$ kHz, however, there are some deviations, as will be shown later in Fig. 3.27.

3.3.2.2. Simple resonator

In practice, we don't know the anyway fictional values of modal mass, modal stiffness, and modal damping, but rather use sound parameters for setting the resonant filters of the synthesis model.

Based on the pole radius

$$R = e^{(-\pi \frac{B}{F_s})} \quad (3.129)$$

and bandwidth B , the coefficients are formulated to equal

$$\begin{aligned} a_2 = R^2 , \quad a_1 = \frac{-4a_2}{1 + a_2} \cos\left(\frac{2\pi f_r}{F_s}\right) , \\ b_0 = (1 - a_2) \sqrt{1 - \frac{a_1^2}{4a_2}} , \end{aligned} \quad (3.130)$$

and $b_1 = b_2 = 0$, leading to the simple resonator (Pirkle 2019, pp. 258–259). It is normalized to unity peak gain and therefore needs to be set to frequency f_r directly and scaled via

$$g = Q \frac{\omega_0}{\omega_d} \quad (3.131)$$

to match the peak gain of the physical system. Contrary to the difference approximation, even at audio rate, the frequency response of the simple resonator is almost identical to that of the original system.

3.3.2.3. Smith-Angell resonator

In a real-time application for auditory augmentation, we seek a resonator that suppresses as much of the direct signal as possible. In addition, we will need matched peaking filters which are able to zero out the resonances. Therefore we seek a resonator that has a symmetric frequency response around its peak. Both requirements get fulfilled by simply adding two zeros at 0 Hz and $F_s/2$, which leads to the Smith-Angell resonator (see Smith and Angell 1982 and also Pirkle 2019, pp. 260–261) with the coefficients

$$a_1 = -2R \cos\left(\frac{\omega_d \omega_r}{\omega_0 F_s}\right), \quad a_2 = R^2, \quad (3.132)$$

$$b_0 = 1 - R, \quad \text{and} \quad b_2 = -b_0 R,$$

with pole radius R as in the simple resonator, $b_1 = 0$, and peak gain scaled to match the original system via

$$g = \frac{\omega_0}{\omega_d}. \quad (3.133)$$

Figure 3.27 shows the three different implementations of the driven harmonic oscillator in comparison to the original system from Fig. 3.5. Note that Q in this example is incredibly low in comparison to the Q -factors that are used in the actual sound model. For higher Q -factors, the differences between ω_0 , ω_d , and ω_r vanish just as well as those between the different implementations.

3.3.2.4. Validation

Czuka (2021) validated the results of the above physical sound model with measurements of real physical plates. A general problem was that the true material parameters of the real plate were not entirely known. While both sounded differently, we assume that human listeners are not able to discriminate the synthetic from a real plate in an absolute identification task (see also Sec. 2.2.16). For an exemplary aluminum plate, spectrogram and magnitude spectrum are shown in Fig. 3.28, separately for synthetic and real plate.

Bibliography

Brémaud, Iris (Jan. 2012a). “Acoustical properties of wood in string instruments soundboards and tuned idiophones: Biological and cultural diversity”. In: *The Journal of the Acoustical Society of America* 131.1, pp. 807–818. DOI: 10.1121/1.3651233.

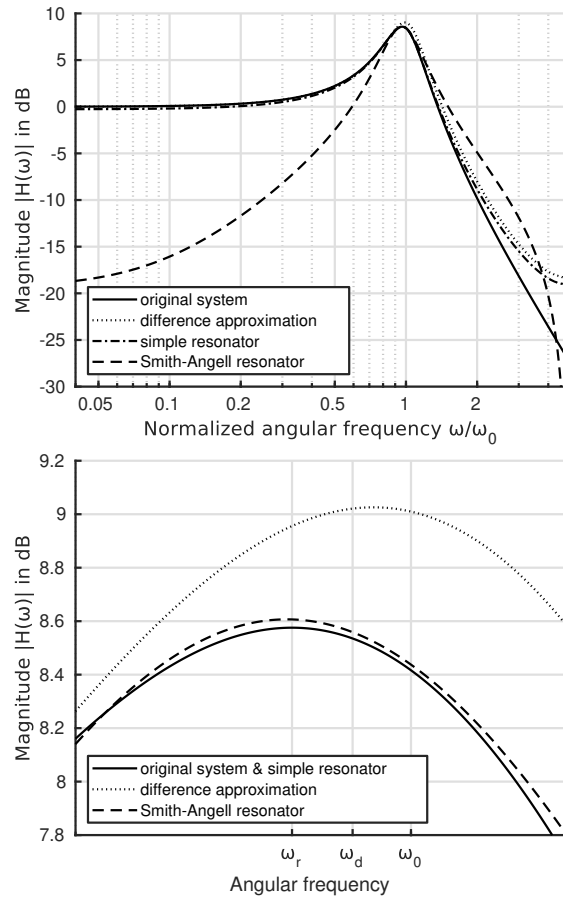


Figure 3.27.: Comparison between 3 digital resonator implementations at audio rate $F_s = 48$ kHz, based on the simple system from Fig. 3.5: a finite difference approximation set with physical parameters and both simple and Smith-Angell resonator set with sound parameters.

- (2012b). “What do we know on ‘resonance wood’ properties? Selective review and ongoing research”. In: *Acoustics Conference*. Nantes, France: Société Française d’Acoustique, pp. 2759–2764.
- Brémaud, Iris et al. (Jan. 1, 2013). “Changes in viscoelastic vibrational properties between compression and normal wood: roles of microfibril angle and of lignin”. In: *Holzforschung* 67.1, pp. 75–85. DOI: 10.1515/hf-2011-0186.
- Chaigne, Antoine and Vincent Doutaut (Jan. 1997). “Numerical simulations of xylophones. I. Time-domain modeling of the vibrating bars”. In: *The Journal of the Acoustical Society of America* 101.1, pp. 539–557. DOI: 10.1121/1.418117.
- Chaigne, Antoine and Christophe Lambourg (Apr. 2001). “Time-domain simulation of damped im-

3. Physical modeling for plausible auditory augmentation

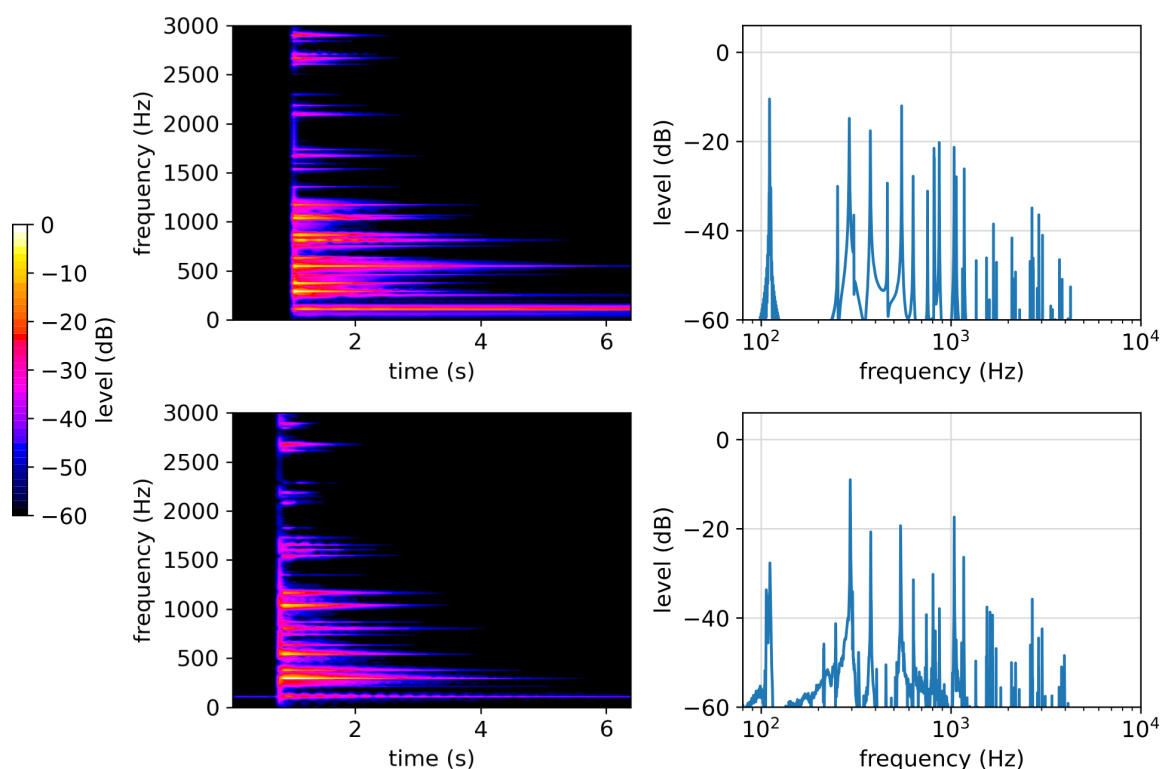
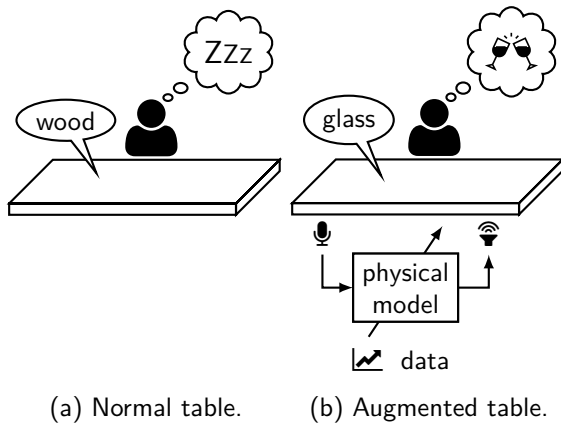


Figure 3.28.: Spectrograms (left) and magnitude spectra (right) of synthesized (top) and recorded (bottom) sound of an impacted aluminum plate. Illustration by Czuka (2021).

- impacted plates. I. Theory and experiments". In: *The Journal of the Acoustical Society of America* 109.4, pp. 1422–1432. DOI: 10.1121/1.1354200.
- Cremer, Lothar, M. Heckl, and B. A. T. Petersson (2005). *Structure-borne sound: structural vibrations and sound radiation at audio frequencies*. 3rd ed. Springer. ISBN: 978-3-540-22696-3.
- Czuka, Martin (2021). "Sound Synthesis and Acoustic Characterization of Rectangular Plates". Master's thesis. University of Music and Performing Arts, Graz, Austria.
- Czuka, Martin, Marian Weger und Robert Höldrich (2021). „Klangsynthese und akustische Erkennung rechteckiger Platten“. In: DAGA - Jahrestagung für Akustik. Vienna, Austria.
- Fletcher, Neville H. and Thomas D. Rossing (2010). *The physics of musical instruments*. 2nd ed. New York, NY: Springer. ISBN: 978-1-4419-3120-7.
- Gutierrez-Lemini, Danton (2014). *Engineering Viscoelasticity*. Boston, MA: Springer. ISBN: 978-1-4614-8138-6. DOI: 10.1007/978-1-4614-8139-3.
- Lutfi, Robert A. et al. (July 2005). "Classification and identification of recorded and synthesized impact sounds by practiced listeners, musicians, and nonmusicians". In: *The Journal of the Acoustical Society of America* 118.1, pp. 393–404. DOI: 10.1121/1.1931867.
- McAdams, Stephen, Antoine Chaigne, and Vincent Roussarie (Mar. 2004). "The psychomechanics of simulated sound sources: Material properties of impacted bars". In: *The Journal of the Acoustical Society of America* 115.3, pp. 1306–1320. DOI: 10.1121/1.1645855.
- McAdams, Stephen et al. (2010). "The psychomechanics of simulated sound sources: Material properties of impacted thin plates". In: *The Journal of the Acoustical Society of America* 128.3. DOI: 10.1121/1.3466867.
- McIntyre, M.E. and J. Woodhouse (June 1988). "On measuring the elastic and damping constants of orthotropic sheet materials". In: *Acta Metallurgica* 36.6, pp. 1397–1416. DOI: 10.1016/0001-6160(88)90209-X.
- Morozov, E V and V V Vasiliev (2003). "Determination of the shear modulus of orthotropic materials from off-axis tension tests". In: *Composite Structures*.

- Ono, Teruaki and Misato Norimoto (Apr. 20, 1983). "Study on Young's Modulus and Internal Friction of Wood in Relation to the Evaluation of Wood for Musical Instruments". In: *Japanese Journal of Applied Physics* 22 (Part 1, No. 4), pp. 611–614. DOI: 10.1143/JJAP.22.611.
- Pirkle, Will C. (2019). *Designing Audio Effect Plugins in C++*. 2nd ed. Routledge. ISBN: 978-0-429-49024-8.
- Putra, A. and D.J. Thompson (Dec. 2010). "Sound radiation from rectangular baffled and unbaffled plates". In: *Applied Acoustics* 71.12, pp. 1113–1125. DOI: 10.1016/j.apacoust.2010.06.009.
- Rossing, Thomas D., ed. (2014). *Springer handbook of acoustics*. 2nd ed. Springer. DOI: 10.1007/978-1-4939-0755-7.
- Schedin, S., C. Lambourg, and A. Chaigne (Apr. 1999). "Transient sound fields from impacted plates: comparison between numerical simulations and experiments". In: *Journal of Sound and Vibration* 221.3, pp. 471–490. DOI: 10.1006/jsvi.1998.2004.
- Smith, Julius O. and James B. Angell (1982). "A Constant-Gain Digital Resonator Tuned by a Single Coefficient". In: *Computer Music Journal*, pp. 36–40.
- Steele, Charles R. and Chad D. Balch (2009). *Introduction to the Theory of Plates*. Division of Mechanics and Computation, Department of Mechanical Engineering, Stanford University.
- Traer, James, Maddie Cusimano, and Josh H. McDermott (2019). "A perceptually inspired generative model of rigid-body contact sounds". In: *International Conference on Digital Audio Effects (DAFx)*. Birmingham, UK.
- Troccaz, Philippe, Roland Woodcock, and Frédéric Laville (Nov. 2000). "Acoustic radiation due to the inelastic impact of a sphere on a rectangular plate". In: *The Journal of the Acoustical Society of America* 108.5, pp. 2197–2202. DOI: 10.1121/1.1312358.
- Warburton, G. B. (June 1954). "The Vibration of Rectangular Plates". In: *Proceedings of the Institution of Mechanical Engineers* 168.1, pp. 371–384. DOI: 10.1243/PIME_PROC_1954_168_040_02.
- Zoghaib, Lionel and Pierre-Olivier Mattei (Aug. 2015). "Damping analysis of a free aluminum plate". In: *Journal of Vibration and Control* 21.11, pp. 2083–2098. DOI: 10.1177/1077546313507098.

4. “AltAR/table”: an experimental platform for plausible auditory augmentation



(a) Normal table. (b) Augmented table.

Figure 4.1.: The concept of AltAR/table.



A condensed version of this chapter in the form of an article is provided by Weger, Hermann, and Höldrich (2022).

In the previous chapter, we derived a physical model of rectangular plates. It is based on subtractive modal synthesis and thus perfectly suited for auditory augmentation. In this chapter, we present AltAR/table¹, an apparatus which employs the model plate in order to plausibly augment a physical interface plate. The basic principle of such an auditory augmentation has already been discussed in Sec. 1.6. Different to Bovermann et al. (2010), we want to avoid visible loudspeakers as far as possible and thus use structure-borne exciters instead, which transform the interface plate into a bending wave loudspeaker. In addition, the target application as augmented table requires a rather large interface plate. Multiple contact microphones and exciters are therefore necessary to allow interaction with the whole surface. The general concept of AltAR/table is once more visualized in Fig. 4.1.

As a start, we assume only one input (contact microphone) and one output (structure-borne exciter). A basic auditory augmentation system then follows the block diagram that is shown in Fig. 4.2. The excitation signal $e(t)$ that is generated by the user’s

¹AltAR stands for alternative auditory reality

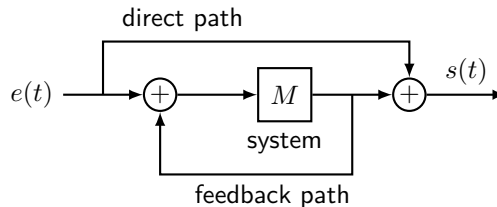


Figure 4.2.: Simplified block diagram of an auditory augmentation system with structure-borne excitation $e(t)$, transfer function M , and airborne output $s(t)$.

action is filtered by the transfer function M of the auditory augmentation system. What reaches the user’s ears is the output signal $s(t)$. In addition to the main signal path, there are two additional paths that are generally undesirable. As both piezoelectric contact microphones and structure-borne exciters are mounted on the same physical structure (plate), some acoustic feedback is introduced that may lead to instability and thus howling. In addition, the original unfiltered excitation signal reaches the ears through the air via the direct path.

The following sections of this chapter describe the different stages of the AltAR/table platform. We start with the hardware apparatus in Sec. 4.1, derive a method to equalize it in Sec. 4.2, and perform actual measurements and calibrations in Sec. 4.3. Both data and audio signal processing are mainly implemented in the SuperCollider 3 language². Sound spatialization, the implementation of the actual sound synthesis model, as well as noise and feedback control are described in Sec. 4.4, 4.5, and 4.6, respectively. Finally, we will examine the technical setup for position tracking in Sec. 4.7 and draw some conclusions in Sec. 4.8.

4.1. Hardware platform

AltAR/table is designed to be put on top of a normal table, to work somehow similar to a touchscreen —

²SuperCollider: <https://supercollider.github.io/>

4. “AltAR/table”: an experimental platform for plausible auditory augmentation

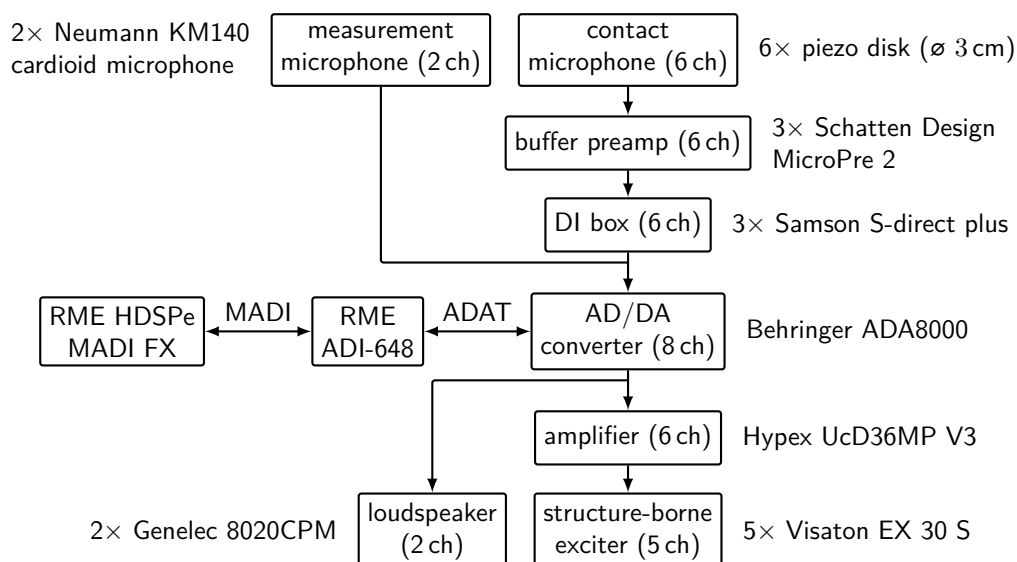


Figure 4.3.: Block diagram of the hardware signal flow of the AltAR/table platform. Unless labeled differently, arrows describe analog cable-connections.

with an auditory display instead of a visual display. It is more a proof of concept than a final product; for usage as “percussion simulator” in psychoacoustic experiments or in teaching, for sound installations, and for evaluation of this kind of auditory augmentation. It therefore strongly draws on professional audio equipment that may be replaced by more embedded and miniaturized technology in the future.

The block diagram of the hardware platform is depicted in Fig. 4.3, while Fig. 4.4 shows a photo. The diagram includes measurement microphones which are only used for calibration. The interface of the hardware platform is based around a 0.7 m × 0.5 m Dibond aluminum-polyethylene-composite plate of thickness 3 mm (Alcan Composites 2006).

The ideal interface plate would be maximally damped and maximally solid so that the excitation signal is captured by the contact microphones but not radiated as airborne sound. At the same time, the interface should be thin and elastic so that the exciters are capable of inducing bending waves which are then radiated as airborne sound. The composite plate represents a compromise between those conflicting interests.

The interface plate sits on top of a frame made of glued particle board (see Fig. 4.5). The region where it overlaps with the frame is covered with felt on both sides, top and bottom, so that it is damped as much as possible. The inner region of the interface plate which can freely vibrate equals the size of a DIN A2 paper sheet, being 594 mm × 420 mm with

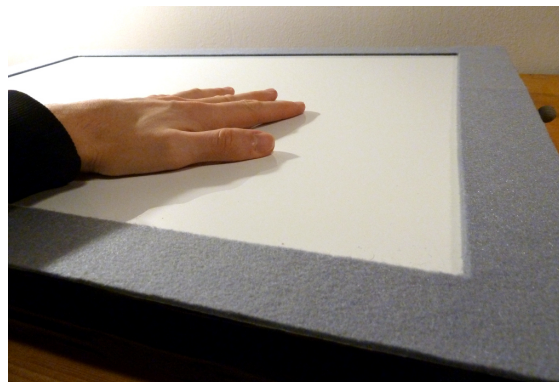


Figure 4.4.: The AltAR/table hardware platform (top).

an area of 0.25 m². The frame itself sits on rubber feet in the corners.

On the bottom side, $J = 5$ structure-borne exciters (Visaton EX 30 S, 8 Ω) and $I = 6$ piezo-electric contact microphones (generic 3 cm piezo disks) are mounted to the interface plate by using 3M VHB double-sided adhesive tape. This kind of adhesive is recommended by Visaton for mounting the EX-series exciters. For the piezo disks, we tested different mounting options (Tesa universal double-sided tape, 3M VHB, modeling clay) and found no significant differences for the given application.

Four contact microphones are placed in the corners, with 5 mm space to the frame, two others

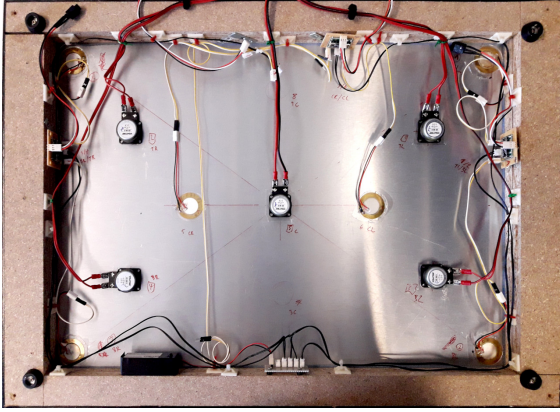


Figure 4.5.: The AltAR/table hardware platform (bottom).

are placed in the center so that their distance between each other (24 cm) equals their distance to the nearby corner-microphones, respectively. The structure-borne exciters are placed halfway between neighboring microphones. This setup ensures a symmetric arrangement with a maximum distance between exciters and piezos.

The piezo disks require a buffer preamp for impedance matching. Three Schatten Design MicroPre 2-channel buffers are attached to the main frame via magnets in order to facilitate their relocation to other prototypes which share the same components. They are powered by either a single 9V battery or a Boss PSA-230P power supply. Their outputs are fed into three Samson S-direct Plus 2-channel DI boxes in order to provide a balanced and symmetric signal for the longer cable run to the Behringer ADA8000 AD/DA converter. The remaining two input channels are designated for two additional microphones that are used for measurements and calibration.

Two outputs of the AD/DA converter are connected to two Genelec 8020CPM monitor speakers which are mounted on the table with Manfrotto Super Clamps and serve as low-frequency extension. The remaining six outputs are directly connected to a Hypex UcD36MP V3 6-channel power amplifier which drives the structure-borne exciters.

The AD/DA converter is connected to an RME HDSPe MADI FX soundcard through an RME ADI-648 ADAT-to-MADI format converter. Audio processing is done on a PC running Debian GNU/Linux.

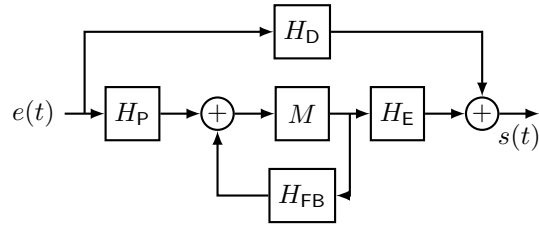


Figure 4.6.: Block diagram including transfer functions of the signal paths.

4.2. Equalization

Until now, the apparatus described in the previous section was assumed to be a transparent sound system that captures the raw excitation signal, e.g., from hand interaction with the plate, and plays back the augmented auditory feedback through a bending wave loudspeaker. In practice, however, piezos, exciters, A/D- and D/A-converters, as well as the physical structure (plate) are far from being transparent. What is captured by the contact microphones is the convolution of the excitation signal with the impulse response of the interface plate. In addition, this transfer function differs not only for every microphone, but also for every excitation position on the plate. Similarly, the bending wave loudspeaker that is driven by the exciters has no flat frequency response at all, but is again filtered by the transfer function and radiation pattern of the plate. To overcome these distortions, all inputs and outputs are individually calibrated by equalization filters.

The simple theoretical block diagram of Fig. 4.2 is therefore extended accordingly (see Fig. 4.6). The excitation signal $e(t)$ is filtered within the structure of the interface plate by transfer function H_P before reaching the contact microphone (in fact this covers also the rest of the input signal path including A/D conversion). The model M is fed with input signal $u(t)$ and outputs $v(t)$. The signal path from D/A conversion, via the exciters and the interface plate to an air microphone is summarized by H_E . The feedback path from $v(t)$ via the exciters back into $u(t)$ is summarized in H_{FB} .

We use the following nomenclature for mathematical description. A system is equally defined by its transfer function $H(s)$, its impulse response $h(t)$, or its spectrum (the Fourier transform of the impulse response) $H(j\omega)$. For better readability and to avoid confusion with subscript indices of piezos ($i = \{1, \dots, I\}$), exciters ($j = \{1, \dots, J\}$), and

4. "AltAR/table": an experimental platform for plausible auditory augmentation

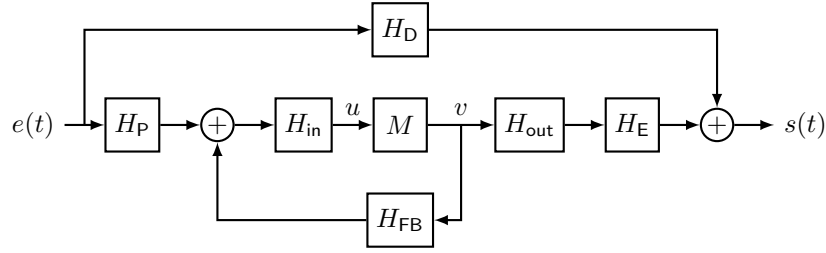


Figure 4.7.: Block diagram including equalization.

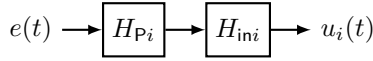


Figure 4.8.: Block diagram for input filtering.

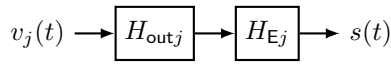


Figure 4.9.: Block diagram for output filtering.

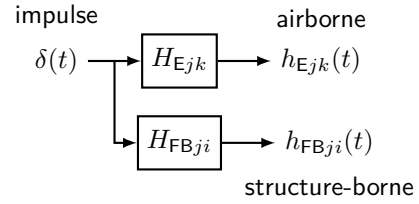


Figure 4.10.: Block diagram for impulse response measurements.

measurement microphones ($k = \{1, \dots, K\}$), H is used interchangeably for transfer function and spectrum without (s) or ($j\omega$); also the time-dependency of the impulse response is omitted.

First we want to reconstruct the true excitation signal $e(t)$ from the input signal $u(t)$ at the microphone. Therefore, we introduce an equalization filter H_{in} which completely cancels the effects of the signal path from excitation to A/D conversion. Its transfer function should therefore be:

$$H_{in} = \frac{1}{H_P} . \quad (4.1)$$

Similarly, we want to hear the true output signal $v(t)$ instead of $s(t)$ at our ears. This is done with the output equalization filter H_{out} :

$$H_{out} = \frac{1}{H_E} . \quad (4.2)$$

The resulting block diagram is shown in Fig. 4.7.

If the direct path is neglected and the feedback-path is assumed to be completely suppressed by feedback control, the input signal $u(t)$ is equal to the true excitation signal $e(t)$ (see Fig. 4.8). Similarly, the signal at our ears $s(t)$ is equal to the output signal $v(t)$ (see Fig. 4.9).

Both types of filters (input and output) are based on impulse response measurements, carried out by the swept-sine method (Farina 2000). For airborne sound measurements, the two Neumann KM140 cardioid microphones (see block diagram in Fig. 4.3)

are placed approximately at the ear positions of a hypothetical user sitting in front of the apparatus similar to sitting on a normal table. The microphones were directed towards the center of the plate. Even if the actual measurement procedure includes several different steps, it is conceptually similar to a perfect impulse being played through the unknown system, and directly delivering the system's impulse response at the receiver.

The desired equalization filter for a given transfer function is actually its inverse. For real-time application, we want a causal filter with minimum group delay. For that purpose, a minimum-phase filter is designed based only on the inverted magnitude response, ignoring the phase.

$h_{E_{jk}}$ represents the impulse response of the system with transfer function or spectrum $H_{E_{jk}}$, from exciter j to measurement microphone k . Similarly, $H_{FB_{ji}}$ represents both transfer function and spectrum of the system between exciter j and contact microphone i . The conceptual impulse response measurement is depicted in Fig. 4.10.

For equalization of exciters, the raw magnitude spectra $|H_{E_{jk}}|$ of the $K = 2$ measurement microphones are energetically averaged to form an overall magnitude spectrum $|H_{E_j}|$ that is approximately valid for all ear positions:

$$|H_{E_j}| \approx \sqrt{\frac{1}{K} \sum_{k=1}^K |H_{E_{jk}}|^2} . \quad (4.3)$$

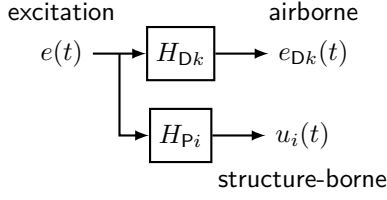


Figure 4.11.: Block diagram for the excitation signal, including direct path and input path of the system.

The desired inverse filter H_{outj} is then

$$H_{outj} = \text{MP} \left\{ (|H_{Ej}| + \rho)^{-1} \right\} \quad (4.4)$$

where ρ is a small regularization parameter and MP is the operation of creating a minimum-phase filter from any given magnitude spectrum (see the end of this section for details). MP includes further regularization.

The input filters H_{ini} do not only depend on the individual contact microphone i and its position, but also on the excitation position (x_0, y_0) to the output of the individual piezo i . If the J structure-borne exciters themselves (without the plate) are assumed to exhibit an approximately flat magnitude spectrum, the impulse response measurements can also be interpreted as the responses to ideal excitations (with $e(t) = \delta(t)$) at the positions of the exciters:

$$|H_{Pi}| \approx \sqrt{\frac{1}{J} \sum_{j=1}^J |H_{FBji}|^2} , \quad (4.5)$$

$$H_{ini} \approx \text{MP} \left\{ (|H_{Pi}| + \rho)^{-1} \right\} . \quad (4.6)$$

It may also be interesting to reconstruct the original airborne signal, i.e., the direct sound $e_D(t)$, based on the input signal $u(t)$ of the piezos. The block diagram in Fig. 4.11 now includes the direct path H_D from the excitation (via the excited interface plate) to the ears.

The desired equalization filter is then the backward path from $u_i(t)$ to $e_{Dk}(t)$. In principle (omitting the channel indices for clarity), the desired inverse filter H_{ex} becomes

$$H_{ex} = \frac{H_D}{H_P} . \quad (4.7)$$

Its impulse response h_{ex} can be convolved with the input signal $u(t)$ to obtain $e_D(t)$ from $u(t)$ directly:

$$e_D(t) = h_{ex} * u(t) \quad (4.8)$$

where $*$ represents the convolution.

We know neither the transfer function H_{Dk} nor the impulse response h_{Dk} of the direct path. However, as we anyway compute a ratio between two spectra in Eq. 4.7, the actual spectra are irrelevant. This fact gets clearer, if the time-domain signals in Fig. 4.11 are expressed by their Fourier transforms in frequency domain,

$$E_{Dk} = H_{Dk} \cdot E_k \quad (4.9)$$

for airborne, and

$$U_i = H_{Pi} \cdot U_i \quad (4.10)$$

for structure-borne sound. The average magnitude spectrum $|\bar{E}_D|$ of the direct signal approximates the combination of the true direct signal E with the average transfer function \bar{H}_D of the direct path:

$$|\bar{E}_D| = \sqrt{\frac{1}{K} \sum_{k=1}^K |E_{Dk}|^2} \approx |\bar{H}_D| \cdot |E| . \quad (4.11)$$

A similar approximation applies for the average structure-borne signal $|\bar{U}|$ that is captured by the piezos:

$$|\bar{U}| = \sqrt{\frac{1}{I} \sum_{i=1}^I |U_i|^2} \approx |\bar{H}_P| \cdot |E| . \quad (4.12)$$

Inserting Eq. 4.11 and 4.12 in Eq. 4.7 eliminates the unknown true excitation signal E and yields

$$H_{ex} \approx \text{MP} \left\{ \frac{|\bar{E}_D|}{|\bar{U}| + \rho} \right\} \quad (4.13)$$

for the desired minimum-phase filter. Instead of impulses or sine sweeps, the measurement is based on a recording of hand-interaction at random locations across the interface plate. The spectra E_{Dk} and U_i are then given by the long-term average spectrum of the recording.

If the excitation signal consists only of impacts with a small hammer, it can be interpreted as a perfect impulse. According to the block diagram in Fig. 4.11, the transfer function H_{Pi} of the input in that case simplifies to

$$H_{Pi} \approx \text{MP} \{ |U_i| \} . \quad (4.14)$$

4. “AltAR/table”: an experimental platform for plausible auditory augmentation

There is yet another way of input equalization, on the basis of simple swept-sine measurements. Assuming a flat frequency response of the exciters themselves, their airborne impulse responses can be interpreted as the impulse responses of an excitation of the interface plate at the exciter positions. By averaging over these, according to the block diagram in Fig. 4.10, the desired filter H_{ex} is given as

$$H_{\text{ex}} \approx \text{MP} \{ \bar{H}_{\text{out}} \} = \text{MP} \left\{ \sqrt{\frac{1}{J} \sum_{j=1}^J |H_{\text{out}j}|^2} \right\}. \quad (4.15)$$

The swept-sine methods bear the advantage that they can be carried out in a standardized and automated procedure with no user interaction required. In addition, practice showed that the system worked best and most robust if only outputs are equalized.

Minimum-phase filters are derived from the measured magnitude spectra in Matlab³, based on the method of Smith (2010, pp. 297–303)⁴. What follows below is a description of the operation of creating a minimum-phase filter from an arbitrary magnitude spectrum $|S(j\omega)|$, as depicted previously by $\text{MP}\{|S(j\omega)|\}$. The desired magnitude spectrum $|S(j\omega)|$ may be constructed from a single recording or impulse response via fast Fourier transform (FFT) or a combination of different magnitude spectra:

1. *Low- and high-pass filtering.* Below a minimum frequency f_{min} (e.g., set to 300 Hz for the exciters, according to their lower cutoff frequency, see Fig. 4.12) and above a maximum frequency f_{max} , the magnitude is either faded to zero with -48 dB per octave or set to stay constant at the value of the cutoff frequency. In case of the exciters, this ensures that the inverse filter does not try to boost low frequencies which just don’t exist. The result is the desired magnitude spectrum of the final filter.
2. *Regularization.* The final filter is going to be applied by convolution whose computational effort grows with the length of the filter. The filter length can be shortened by smoothing the spectrum so that less coefficients are needed to achieve it. For that purpose we apply a $1/8$ -octave smoothing on the magnitude spectrum.

3. *Get minimum phase filter.* The final filter must be stable and causal and exhibit a low latency, i.e., group delay. This is achieved by a filter whose poles and zeros are all within the unit circle, i.e., their radius is below 1. Such a filter is said to be minimum phase. The real-valued magnitude spectrum is converted to a complex-valued minimum phase spectrum via the cepstral method which mirrors the coefficients to the inside of the unit circle. (Smith 2010, pp. 297–303)
4. *IFFT.* The inverse FFT yields the impulse response of the filter in time domain.
5. *Cropping and Fading.* The impulse response is cropped at the point where its envelope reaches -60 dB of its maximum value (i.e., at the -60 dB decay time). This is assumed to be sufficient for the given purpose. Finally, a fade-out is applied to the impulse response. It has the shape of half a period of a raised cosine (i.e., the second half of a symmetric Hann window) and spans the last 10% of the impulse response.

4.3. Measurements

For the measurements, the same apparatus was used as described in Sec. 4.1. Measurements were performed in an acoustically treated room of $6.20 \text{ m} \times 4.33 \text{ m} \times 3.43 \text{ m}$ (length \times width \times height) size.

As the room still exhibits some audible resonances, a pair of Neumann KM140 microphones was used for airborne sound measurements; their cardioid directivity pattern is assumed to suppress the influence of the room more than would be the case with omnidirectional measurement microphones.

Impulse response measurements were performed with the exponential sine sweep (ESS) method (Farina 2000). The sine sweep signal was generated in Matlab and saved as a PCM wave-file in 32 bit floating point format. The procedure is mainly based on the Pd⁵ implementation by Vetter and Rosario (2011).

The sweep spans 14 octaves, from 1.35 Hz to $F_s/2$ (F_s is the sampling frequency), with 7s duration; a fade-in is applied during the first octave. This ensures that ripple between 4 Hz and 20 kHz lies below $\pm 0.1 \text{ dB}$ for measured magnitude spectra.

³MathWorks Matlab: <https://mathworks.com>

⁴a digital copy is available online: https://ccrma.stanford.edu/~jos/pasp/Fitting_Filters_Measured_Amplitude.html

⁵Pure Data (Pd): <https://puredata.info>

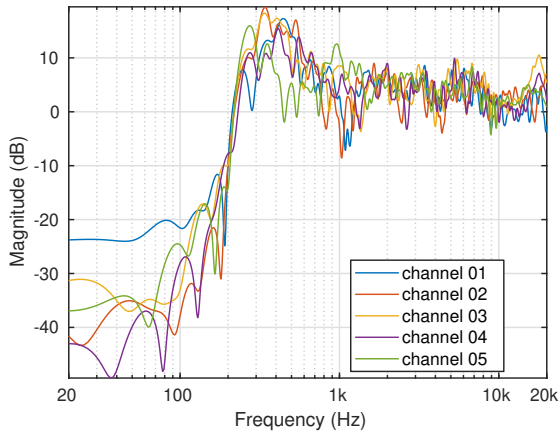


Figure 4.12.: Smoothed magnitude spectra of the structure-borne exciters, measured with airborne microphones (with arbitrary normalization). Channel 05 is the center exciter.

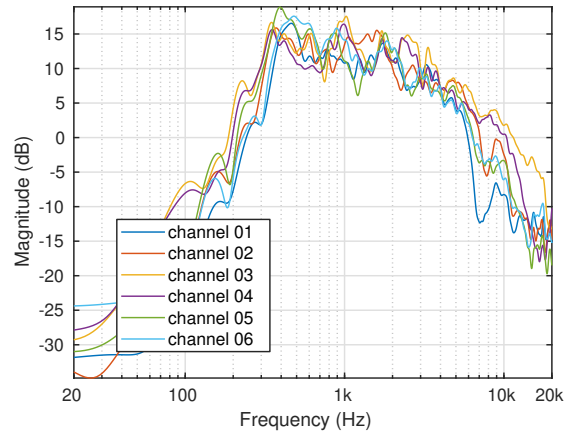


Figure 4.13.: Smoothed magnitude spectra of the piezos, estimated via sine sweeps from exciters (with arbitrary normalization). Plots for the individual piezos, averaged over exciters.

Fade-out was found to be unnecessary. The sweep starts and stops at a zero-crossing. The inverse sweep for deconvolution is a time-reversed version of the original sweep, including additional frequency-dependent amplitude compensation to account for the pink spectrum of the original sweep. Each measurement was performed 5 times for improved SNR.

The measurements were carried out in Pd and processed in Matlab. First, the repetitions are averaged in time-domain. Then the averaged sweep recording is deconvolved with the amplitude-compensated inverse sweep. Only the causal part of the result is taken as impulse response. The impulse responses are normalized to the maximum of all channels within each type (exciter to structure-borne microphone, exciter to airborne microphone, loudspeaker to airborne microphone), so that the maximum equals 0 dBFS.

As a by-product, the maximum of the cross-correlation between the original sweep and its recording yields the round-trip latency of the entire signal chain including D/A- and A/D-conversion.

Magnitude spectra of the signal path between exciters and airborne microphones are shown in Fig. 4.12; between exciters and contact microphones in Fig. 4.13, with 1/24-octave smoothing. Figure 4.14 shows the magnitude spectra of the inverse filters of the structure-borne exciters that are used for equalization. Inverse filters of the contact microphones are shown in Fig. 4.15. The signal path between the 5 exciters and one exemplary contact microphone is shown in Fig. 4.16.

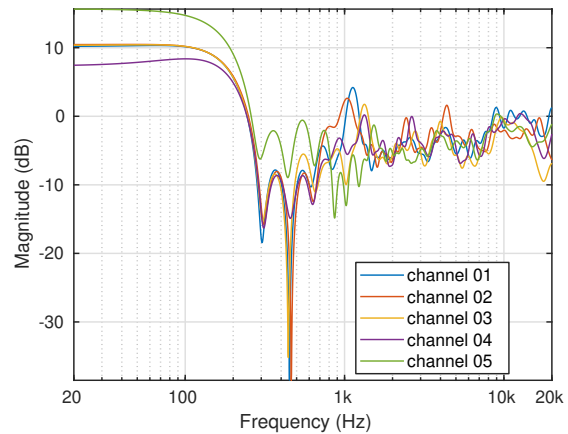


Figure 4.14.: Magnitude spectra of the inverse filters H_{out_j} (with arbitrary normalization), designed to equalize the frequency response of the structure-borne exciters.

The round-trip latency of the system is the time that an analog vibration of the plate takes to pass the complete signal chain including the analog path from the piezo disk to the A/D-converter, the digital block-processing, D/A-conversion, electrical signal path to the structure-borne exciter, and the traveling wave from exciter back to piezo. In theory, this round-trip latency could be directly derived from the measured impulse responses between exciters and piezos. In practice, however, the measurements have been performed with a different software (implemented in Pd) and even different buffer size of the audio driver (for stability and precision instead

4. “AltAR/table”: an experimental platform for plausible auditory augmentation

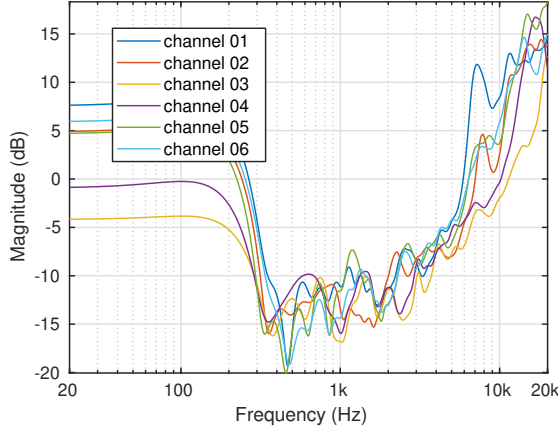


Figure 4.15.: Magnitude spectra of the inverse filters H_{ini} (with arbitrary normalization), designed to equalize the frequency response of the contact microphones.

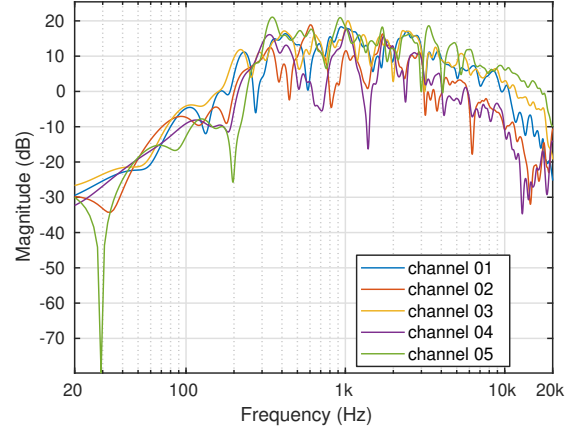


Figure 4.16.: Smoothed magnitude spectra of the signal path between exciters and the 3rd contact microphone (center left), with arbitrary normalization. Channel 05 corresponds to the center exciter.

of low latency), without passing the whole system. The round-trip latency of the whole auditory augmentation system is therefore derived from the optimal delay of the feedback cancellation system (see Sec. 4.6.3). For the longest signal path (longest distance between exciter and piezo) it equals 176 samples or 3.7 ms.

4.4. Spatialization and low-frequency extension

Each input channel, i.e., piezo-electric contact microphone, is processed separately and then spatialized to the output channels, i.e., structure-borne exciters. The positions of piezos and exciters on the interface plate are given in Tab. Tab. 4.1. Spatialization of piezos to exciters is realized via distance-based amplitude panning (DBAP), following its original definition by Lossius et al. (2011).

DBAP delivers individual gain factors based on the distances between the I sources and J speakers. Each distance d_{ij} between the i -th source and the j -th speaker is defined as

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 + r_s^2} \quad (4.16)$$

where (x_i, y_i, z_i) and (x_j, y_j, z_j) are the Cartesian coordinates of source and speaker, respectively. The additional parameter $r_s \geq 0$ m controls the hypothetical size of the source by adding spatial blur.

Table 4.1.: Coordinates of inputs (contact microphones) and outputs (structure-borne exciters), relative to the upper left corner, in mm.

inputs			outputs		
i	x_i	y_i	j	x_j	y_j
1	20	20	1	97	115
2	574	20	2	497	115
3	175	210	3	97	305
4	419	210	4	497	305
5	20	400	5	297	210
6	574	400			

The individual gain for each pair of source and speaker, i.e., the gain of source i for speaker j is given by

$$g_{ij} = d_{ij}^{-a} / k_i \quad (4.17)$$

where a describes the loss per distance

$$a = R/20 \log_{10}(2) \quad (4.18)$$

with roll-off $R \geq 0$ dB which defines the loss per distance doubling in dB. All gains concerning one source i are normalized by

$$k_i = \sqrt{\sum_{j=1}^J d_{ij}^{-2a}} \quad (4.19)$$

to achieve constant sound intensity irrespective to the source position.

For AltAR/table, we use a roll-off $R=6$ dB, and an effectively negligible source size $r_s=1$ mm.

The measurements of the previous section revealed that the exciters are not able to radiate sufficient power at frequencies below 200 Hz. The frequency range is therefore extended down to around 70 Hz by two Genelec 8020CPM loudspeakers. Frequency crossover is set to 250 Hz through 2nd-order low-pass filters for exciter outputs and 2nd-order high-pass filters for loudspeaker outputs. The exciter channels are distributed to the loudspeakers by a simple cosine panning law which preserves constant power.

4.5. Subtractive modal synthesis model

The main purpose of the AltAR platform is to host the physical model of a rectangular plate, i.e., to capture the original auditory feedback, feed it through the model plate, and project the augmented auditory feedback into the physical environment of the user. The physical model itself has already been explained in Ch. 3. We will now briefly describe its application, implemented in the SuperCollider 3 language.

The individual resonances are created via subtractive modal synthesis, based on a bank of parallel Smith-Angell resonators (see Sec. 3.3.2.3), each with an individual gain, frequency, and Q-factor. All resonators are fed with the same input signal; their outputs are summed. The physical model itself can also be regarded as a parameter conversion that maps from physical parameters to intermediate sound parameters that are rendered by the resonators.

The computation of natural frequencies $f_{0,mn}$ and Q-factors Q_{mn} follows the procedure of Sec. 3.2.5 and 3.2.6. Filter frequencies are set so that the peak occurs at ω_r , while both gain and frequency are fine-tuned via the ratio ω_0/ω_d (see Sec. 3.3.2.3). The input gain $G_{in,mn}$ includes the amplitude weights due to indentation hardness (Sec. 3.2.9) as well as excitation position and modal mass (Sec. 3.2.7). The output gain $G_{out,mn}$ describes the radiation efficiency (Sec. 3.2.10). The model parameters are set either by external data or by the graphical user interface shown in Fig. 4.17.

Each time a model parameter is changed, the dependent sound parameters are updated and sent to filters in the real-time synthesis engine that runs

on the SuperCollider server. An exponential lag of 50 ms per 60 dB for all sound parameters ensures smooth transitions. Morphing between model plates of different physical parameters is possible by continuous transitions between the individual parameter values. For some parameters such as boundary conditions, however, it is difficult to morph between two states. In such cases it is possible to convert both models to the sound parameter domain and morph these instead. In both cases, only one instance of the synthesis model is needed. Another morphing strategy is to cross-fade between two or more sound models. For that purpose, AltAR uses different layers to be rendered simultaneously. In addition, different models can be used for different spatial regions of the interface plate, each defined by arbitrary polygons of vertices. These are called zones. The excitation signal is routed to all zones which cover the current excitation position. Every zone may exhibit one or more layers which may consist of one or more models.

The overall impulse response of the parallel resonator filterbank is simply the sum of the impulse responses of the individual resonators. For an exemplary aluminum plate, the overall magnitude spectrum of the resonator filterbank is shown in Fig. 4.18. Destructive additions lead to notches between resonances, depending on the phase responses of the filters. While the exact phase responses and thus notch frequencies differ from those of an actual physical plate, this has only minor impact on the perceived sound. The effect of the notches can be reduced by alternating the sign between the individual signals before summation, as was proposed by Noisternig (2017) for gammatone filterbanks.

4.6. Noise and feedback control

AltAR/table is a rather unstable feedback system. Not only it implements a closed loop between contact microphones and exciters through the same physical plate, but it even inserts about one hundred steep resonant filters within that loop. Without some sort of noise and feedback control, it already blows up by its internal noise, before even touching it.

To suppress noise in irrelevant frequency regions below 80 Hz and above 10 kHz, the input signal is pre-conditioned by 2nd-order high-pass and low-pass filters, respectively. In addition, a simple noise gate is set just above the noise level to make sure that the system stays quiet if unused.

4. "AltAR/table": an experimental platform for plausible auditory augmentation

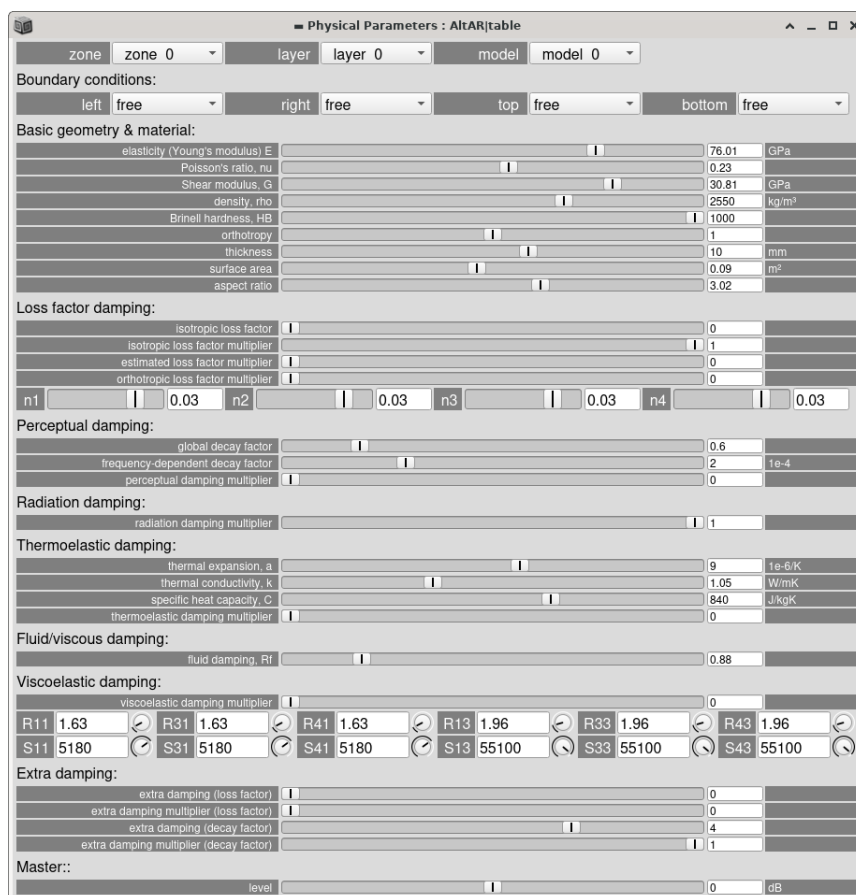


Figure 4.17.: Graphical user interface for the physical parameters of AltAR/table.

The unstable feedback loop is controlled in four different ways. The most prominent source of howling is created by the resonators themselves. These are addresses in Sec. 4.6.1. The three remaining stages comprise frequency shifting (Sec. 4.6.2), feedback cancellation (Sec. 4.6.3), and feedback suppression (Sec. 4.6.4).

4.6.1. Anti-resonator

If we look at the magnitude response of the system in Fig. 4.18, it is pretty obvious that howling is most likely to occur at the resonant frequencies themselves. We therefore seek a filter which reverts the magnitude response of the resonator filterbank at the input, as shown in Fig. 4.19. The analytic inversion of a parallel IIR filterbank is not trivial, and the fact that the coefficients of the individual resonators need to be constantly updated, makes an FIR approximation rather complicated, especially with their long impulse responses. We therefore

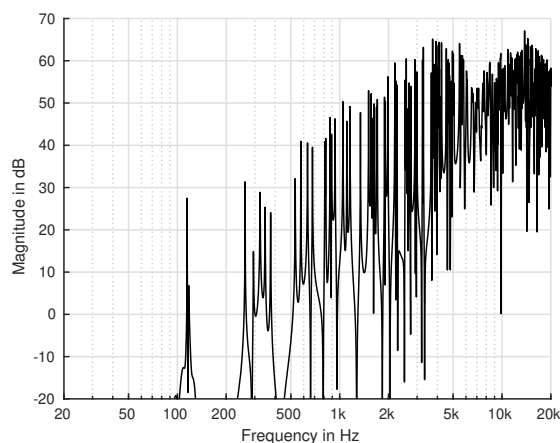


Figure 4.18.: Overall magnitude spectrum of the summed resonators for an aluminum plate.

try a different approach. The parallel resonators are equalized by a bank of serial notch filters, so that in consequence, if both filterbanks are applied

“The world’s most expensive patch cable.”
— R. Höldrich, 2017

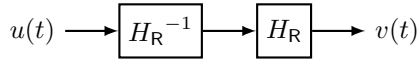


Figure 4.19.: The principle of resonator H_R and anti-resonator H_R^{-1} .

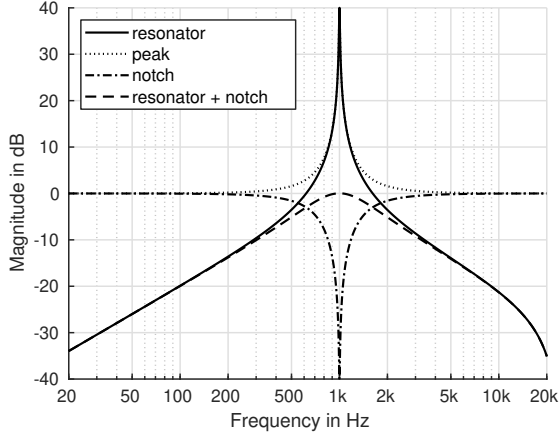


Figure 4.20.: Magnitude responses of a resonator ($f_r=1$ kHz, $Q=100$, gain of 40 dB) together with its matched peak EQ, notch EQ (anti-resonator) and the combination of resonator and notch EQ.

one after another, the magnitude of the resonator filterbank is clipped at a settable threshold level. The desired transfer function of this anti-resonator is therefore

$$H_A = \frac{1}{1 + H_R/g_{RT}} \quad (4.20)$$

where g_{RT} is the threshold gain. The anti-resonator must match the magnitude response of the resonator in the frequency range above the threshold, independent of the actual threshold value. This is achieved by a 2nd-order constant-Q peaking EQ filter (see Pirkle 2019, pp. 279–280). Between positive and negative gains, the magnitude response of the filter is simply mirrored around 0 dB; the gains are set to

$$g_{Amn} = \frac{g_{RT}}{g_{Rmn}} \quad (4.21)$$

where g_{Rmn} is the combined input and output gain of mode m/n which includes all gains from Sec. 3.2.7 except those connected to excitation position—otherwise howling might occur if the excitation position changes after an impact.

Figure 4.20 exemplarily shows the magnitude responses of one resonator, its matched anti-resonator, and their combination. The anti-resonator obviously just makes the resonator obsolete. In Sec. 4.6.2, a frequency shift is therefore introduced between both filterbanks.

While this might work almost perfectly for a single pair of resonator and notch filter, the channel crosstalk between neighboring (parallel) resonators is overestimated by the (serial) notch filters. In case of high modal density, this leads to extreme attenuation of the signal, way down below quantization noise, so that a recovery is impossible. In practice, the Smith-Angell resonators have almost negligible crosstalk, so that their individual peak gains can be assumed to stay equal after summation. We therefore seek the optimal gains of a peaking filterbank, i.e., the gains we need to set so that the true peak gains match those of the resonator filterbank. The reciprocal optimal gains are then the command gains of the notch filterbank. Abel and Berners (2004) proposed to retrieve the optimal gains by solving the system of linear equations that can be created from the magnitude responses of the resonators evaluated at their peak gains (see also Välimäki and Reiss 2016, pp. 14–17). Note that we are only interested in peak gains and assume that the frequency regions in between will automatically be sufficiently matched.

The vector of command gains \mathbf{g}_c describes the peak gains $g_{R,mn}$ in dB, written as a row vector:

$$\mathbf{g}_c = 20 \log \left([g_{R,01} \ g_{R,02} \ \dots \ g_{R,MN}]^T \right) \quad (4.22)$$

with a length equal to the total number of modes $K=M \cdot N$.

The results are even better if the command gains include the crosstalk between resonators. Assuming the worst case of in-phase interference, the true peak gains in dB are then the sum of the K individual magnitude responses evaluated at the observed resonant frequency ω_{rk} :

$$g_{Rk} = 20 \log \prod_{k=1}^K |H_{Rl}(\omega_{rk})| \quad (4.23)$$

The matrix of filter gains \mathbf{B} is constructed by evaluating the magnitude responses of the anti-resonators in dB at all peak frequencies for a gain

4. “AltAR/table”: an experimental platform for plausible auditory augmentation

of 1 dB:

$$\mathbf{B} = 20 \log \begin{bmatrix} |H_{A1}(\omega_{r1})| & |H_{A1}(\omega_{r2})| & \dots \\ \vdots & \ddots & \\ |H_{AK}(\omega_{r1})| & \dots & |H_{AK}(\omega_{rK})| \end{bmatrix} \quad (4.24)$$

with the magnitude response $|H_{Ak}|$ and peak frequency ω_{rk} of anti-resonator k . All elements of the main diagonal are always equal to 1; non-zero elements besides the main diagonal describe overlap, i.e., crosstalk, between filters. For more precise results at high gains, the frequency responses may also be retrieved for an arbitrary value; the gain matrix is then divided by that value.

The optimal gains become

$$\mathbf{g}_{\text{opt}} = \mathbf{B}^{-1} \mathbf{g}_c \quad (4.25)$$

The inverse matrix only needs to be updated if peak frequencies or Q-factors are changed. In case of identical filter frequencies or high bandwidths, \mathbf{B} might become singular and thus non-invertible, which requires additional regularization. In the lack of a proper linear equation solver in SuperCollider, the command peak gains, peak frequencies, and Q-factors are sent via open sound control (OSC)⁶ to a Python script which evaluates the magnitude responses and solves the linear system in a least squares sense.⁷ The resulting optimal gains are then sent back. We already know that the optimal gains must be lower than zero (no amplification), and higher or equal to the command gains (no over-reduction). These bounds are set as constraints to the linear solver; further regularization is then no longer required. A mere matrix inversion would not allow such constraints, making the use of an iterative solver necessary. If parameters are changed constantly, a realistic plate with 100 modes leads to high computational effort for the linear least-squares solver. As the audio signal processing can only run on a single processor core; however, the remaining cores of recent CPUs anyway have idle capacity for that task. Figure 4.21 shows the magnitude response of the anti-resonator filterbank with and without gain optimization.

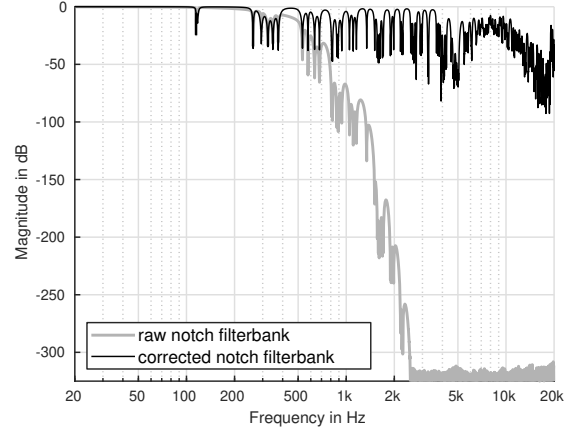


Figure 4.21.: Magnitude response of the optimized notch filterbank for the resonator filterbank shown in Fig. 4.18. The grayed-out spectrum shows the raw notch filterbank without gain optimization.

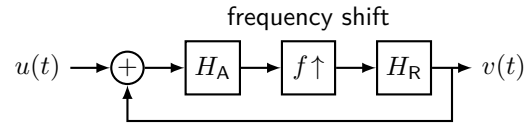


Figure 4.22.: Block diagram including feedback suppression.

4.6.2. Frequency shifting

One strong tool for howling prevention is frequency shifting (see, e.g., Berdahl and Harris 2010). The benefit, however, is at the cost of inharmonic distortion. In our case of broadband impact sounds, this cost is acceptable. We anyway need the frequency shift for an entirely different reason: to move the signal energy that remains after the anti-resonator towards the relevant frequency range of the resonator, according to the block diagram in Fig. 4.22. In case of equidistant resonant frequencies, i.e., a harmonic relationship between partials, a shift by half the base frequency moves the remaining energy straight into the resonant frequencies. However, for the modeled plates, only statistical considerations are possible.

The search for an optimum value requires a second objective to be taken into account. In general, the larger the frequency shift f_{shift} , the larger the maximum stable gain (MSG) (Berdahl and Harris 2010). It was reported that an optimal value for the

⁶open sound control (OSC):

<https://opensoundcontrol.stanford.edu/>

⁷using `lsq_linear` from `scipy.optimize`

frequency shift f_{shift} is half the average frequency distance between two magnitude peaks in the spectrum of the room response (Waterschoot and Moonen 2011). For an average room, this inter-peak difference is about 10 Hz, leading to $f_{\text{shift}} = 5$ Hz, which is almost inaudible for speech and music signals, and may lead to an increased MSG by up to 14 dB (Waterschoot and Moonen 2011). Berdahl and Harris (2010) reported that the increase in MSG was high for cardioid microphones (3.5 dB for $f_{\text{shift}} = 8$ Hz), but low for hearing aids (2 dB for $f_{\text{shift}} = 30$ Hz). In case of bars and plates, the difference between resonances can easily reach 200 Hz, which confirms our informal tests which led to an optimal $f_{\text{shift}} \approx 100$ Hz—still barely audible for the broadband excitation signal of the impact.

The frequency shift is performed via single side-band modulation (SSB), where all frequencies are shifted by a fixed amount.⁸ Pitch shifting instead of frequency shifting would preserve the harmonic relationships between partials; however, low frequencies wouldn't be sufficiently shifted, thus limiting the performance of feedback suppression (Berdahl and Harris 2010).

For SSB, instead of an actual Hilbert transform, the input signal is fed into a pair of all-pass filters whose outputs are approximately 90° out of phase (both different from the original), and consequently interpreted as the real and imaginary part of a complex signal. A (complex) multiplication with a (complex) sinusoid with frequency f_{shift} leads to a frequency shift of the input signal by exactly that amount. If $H_R(f)$ describes the complete transfer function of the system without frequency shift, the transfer function including the frequency shift becomes

$$H_{R,f\uparrow}(f) = \prod_{n=1}^{+\infty} H_R(f + n f_{\text{shift}}) . \quad (4.26)$$

Wherever the magnitude $|H|$ of the overall system is greater than 0 dB, the output grows with each cycle around the loop, leading to unstable behavior. The frequency shift spreads the energy across frequencies and thus reduces such peaks so that they may be pulled below 0 dB, hence stabilizing system.

4.6.3. Feedback cancellation

Due to the strong coupling between contact microphones and structure-borne exciters through the

⁸for real-time implementations see `FreqShift` in `SuperCollider` or example `H09.ssb.modulation.pd` in `Pd`.

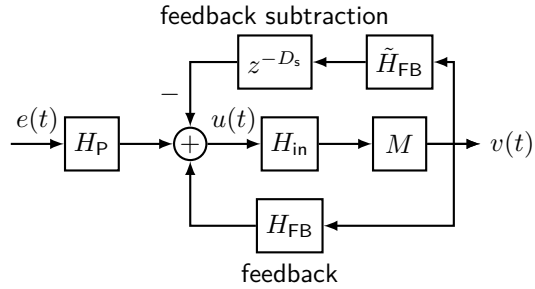


Figure 4.23.: Block diagram of the feedback cancellation.

interface plate, simple feedback suppression is not enough. Instead of just fixing the symptoms of the existing feedback, it would be better to completely prevent it from happening. For that purpose, the feedback path is subtracted from the input signal $u(t)$, in order to reconstruct the true excitation signal $e(t)$ without feedback. The general concept of this form of feedback cancellation is depicted in Fig. 4.23.

The hardware can be assumed to be time-invariant, so that a single measurement of H_{FB} is sufficient. Note that this is actually a MIMO (multiple input, multiple output) system with J outputs (exciters) and I inputs (contact microphones), so that J sweeps are necessary to obtain the $I \times J$ impulse responses. The required measurements have already been carried out in Sec. 4.3 for equalization, and can be re-used for the additional purpose.

The beginning of the measured impulse responses $h_{FB,ji}$ is cut at 32 samples before the shortest signal path delay, so that individual latency differences remain, but additional time is gained for the partitioned convolution. The remainder is truncated at 1024 samples; the first 16 samples are faded in and the last 10% are faded out with a raised-cosine envelope.

As the measurements are performed with a different software setup than the auditory augmentation (Pd instead of SuperCollider), the measured round-trip latencies are not transferable. A rough guess is based on the difference in block size as well as the partition size used for partitioned convolution with the measured impulse responses. The global time delay D_s and subtraction gain g_s are fine-tuned by hand so that $u(t)$ is minimized if pink noise is sent to all outputs at once. The feedback cancellation can be adjusted to affect only frequencies up to a specific frequency by including a low-pass filter in the feedback cancellation path. This filter,

4. “AltAR/table”: an experimental platform for plausible auditory augmentation

however, needs to be ideal zero-phase (i.e., ideal), hence non-causal, and is therefore already applied within pre-processing in Matlab⁹. In the end, the feedback cancellation proved stable and effective over the whole frequency range, so that filtering was not even necessary.

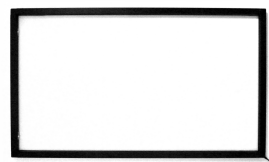


Figure 4.24.: Optical multi-touch frame.

4.6.4. Notch filters

Finally, after activation of the complete system including equalization filters, resonators, anti-resonators, frequency shift, and feedback cancellation, additional howling frequencies can be found in the way a sound engineer would do it: by sending pink noise to the outputs and raising the level until howling occurs. As soon as an additional notch filter at the howling frequency is inserted into the signal path, the process is repeated to find more howling frequencies. The system is ready to use if the noise signal can be played back at an adequate level without howling.

4.7. Tracking of contact position and hand damping

From everyday life, we know already that any physical object usually sounds different at different excitation positions. In musical instruments, we take advantage of this behavior by deliberate control of the plucking or striking position for musical expression. In Sec. 3.2.7, we learned that this location-dependent timbre is evoked by the shapes and thus amplitudes of the individual modes at the excitation position. For realistic and thus plausible auditory augmentation, this sonic behavior needs to be taken into account. For that purpose, the excitation position, i.e., the tip of the finger, mallet, or pen, is tracked in order to pass the coordinates to the physical model in real time via OSC. At different stages during the development of AltAR/table, we tried different tracking techniques. While marker-based infrared tracking offers the highest precision, the less professional solutions don't require such cumbersome infrared-reflective markers.

In addition, some of the solutions make it possible to detect damping by the flat hand, which raises the amount of realism to the next level.

4.7.1. Markerless position tracking

It would be perfect, if position tracking could be done precisely without being noticed by the user. With infrared cameras that achieve motion sensing through structured light or time of flight calculations, such position tracking is possible. One of the first pilot studies of the augmented table, the Auditory Coloring Book (described in detail in Sec. 8.1.2) makes use of this technology (see also Weger et al. 2018). For that purpose, markerless optical tracking was realized with the Microsoft Kinect v2 sensor. The actual identification of the tip of the pen was done in Processing with KinectPV2¹⁰ and OpenCV¹¹ libraries. Based on the results of the pilot study, however, we came to the conclusion that a more robust and accurate solution is necessary.

For our 2D interface plate, we do not necessarily need the height information. A 2D position tracking similar to a touch screen is already enough. We therefore use a NECO 32 inch IR multi-touch overlay frame (see Fig. 4.24) that allows simultaneous tracking of up to 10 individual points ≥ 5 mm in size. The frame itself is constructed from aluminum bars of 19 mm width and 8.7 mm thickness, with effective tracked region of 702.1 mm \times 395.3 mm, and is usually intended to be attached on top of ordinary TV or computer screens. It is connected via USB 2 and delivers HID-compatible data encoded in the multi-touch protocol¹² which includes positions as well as sizes of the tracked fingers (or any other objects) as soon as they are located in the tracked space within the frame. A Python script is used to grab the device in order to prevent the operating system from instantly using it, and to transform the raw pixel coordinates into normalized coordinates (between 0 and 1 in both directions) and forward

⁹via the `filtfilt` function

¹⁰KinectPV2 for Processing:

<https://github.com/ThomasLengeling/KinectPV2>

¹¹OpenCV for Processing:

<https://github.com/atduskgreg/opencv-processing>

¹²Multi-touch (MT) protocol:

e.g., <https://www.kernel.org/doc/html/latest/input/multi-touch-protocol.html>

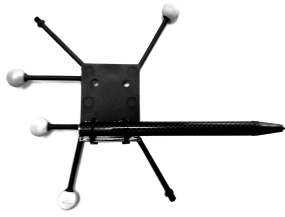


Figure 4.25.: Pencil with infrared-reflective markers.

</> it to SuperCollider via OSC. The source code is available in Source 4.1.

The position of the most recently detected point is taken as the current excitation position. For each point, the major and minor axis of an assumed elliptical shape are included in the multi-touch protocol. Based on the physical frame dimensions, we are therefore able to retrieve the area of each ellipse as an estimate for the point size. The precision is not sufficient to distinguish the type of interaction (e.g., tip of the finger vs. fingernail vs. pen), but suffices for detecting the heel or palm of the hand. The sizes of all detected points are summed; the result is then mapped to the additional damping, simulating continuous hand contact. This enables users to dampen the model plate in a realistic manner while interacting naturally with the interface plate.

4.7.2. Marker-based position tracking

Highest speed and accuracy is delivered by marker-based infrared tracking, on the cost of reflective markers. We use it only for tool-mediated interaction, namely through a ballpoint pen that is equipped with markers as shown in Fig. 4.25. In case of the Exploration Table prototype in Sec. 7.3.2, a soft mallet was used. Only such prepared tools are able to be tracked; the detection of hand contact is not possible. In the scope of this work we use 8 OptiTrack Flex 13 cameras together with the Motive:Body software running on a dedicated PC. The real-time tracking data is forwarded to OSC by a small wrapper application¹³.

OptiTrack tracks so-called rigid bodies that are constructed of three or more infrared-reflective markers. The absolute position that is sent in Cartesian coordinates is the centroid of all associated markers. While three are sufficient to define position and rotation, a fourth marker is usually added for redundancy. Arranged in a plane, the centroid or center

¹³pyNatNat: <https://git.iem.at/tracking/pyNatNat>, NatNat: <https://git.iem.at/tracking/NatNat>

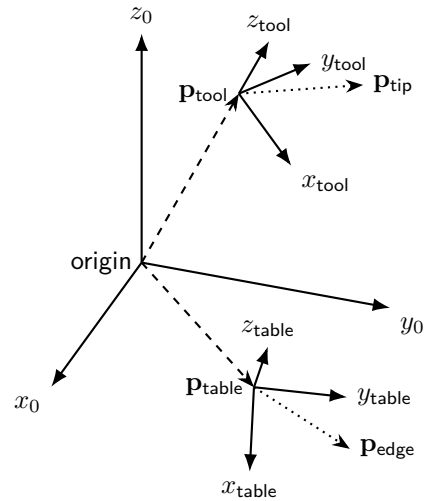


Figure 4.26.: Coordinate transformations of tracked pencil and table.

of mass equals the intersection point of the two bimedians (i.e., the lines joining the center points of two opposing sides). The absolute orientation, relative to the default orientation upon creation, is given in form of a quaternion.

Besides the tool, also the interface plate is equipped with markers to be tracked as rigid-body object. We will denote the absolute center positions of tool and table as \mathbf{p}_{tool} and $\mathbf{p}_{\text{table}}$, while the orientation is rewritten as 3×3 rotation matrix¹⁴ \mathbf{R}_{tool} and $\mathbf{R}_{\text{table}}$ for tool and table, respectively. Figure 4.26 illustrates the positions and rotation of tool and table within the global coordinate system.

In the default orientation, the difference $\Delta \mathbf{p}_{\text{tool}}$ in Cartesian coordinates between the tip of the tool and the centroid of the attached markers is measured. In case of the interface plate, $\Delta \mathbf{p}_{\text{table}}$ describes the difference between the coordinates of the upper left corner and the centroid of all markers. These calibration measurements are carried out by hand.

What we actually want is the excitation position, i.e., the position of the tip of the tool, relative to the upper left corner of the interface plate. The absolute position of the tip is

$$\mathbf{p}_{\text{tip}} = \mathbf{p}_{\text{tool}} + (\mathbf{R}_{\text{tool}} \cdot \Delta \mathbf{p}_{\text{tool}}) . \quad (4.27)$$

This position is further transformed by the table to

¹⁴Conversion algorithms between quaternion and rotation matrix are adopted from the JavaScript 3D library three.js: <https://github.com/mrdoob/three.js>

4. “AltAR/table”: an experimental platform for plausible auditory augmentation

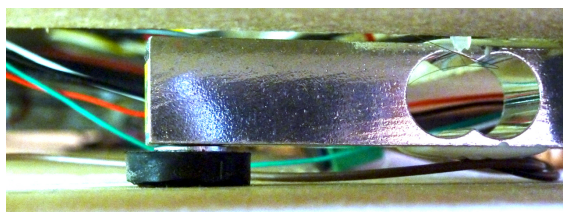


Figure 4.27.: Pressure sensing via load cells.

retrieve the excitation position

$$\mathbf{p}_{\text{ex}} = \mathbf{R}_{\text{table}}^{-1} \cdot (\mathbf{p}_{\text{tip}} - \mathbf{p}_{\text{table}}) + \Delta \mathbf{p}_{\text{table}} \quad (4.28)$$

The resulting coordinate \mathbf{p}_{ex} describes the position (in m) relative to the position and orientation of the upper left corner of the interface plate. The x - and y -axis are finally normalized, i.e., divided by the interface dimensions, to be forwarded to the physical model via OSC, thus controlling the excitation position. The z -axis (height) is kept in physical units; it is used in the experiment in Sec. 5.2 for detecting contact between pen and plate.

4.7.3. Pressure sensing

In order to allow damping by the palm of the hand even in case of marker-based tracking, AltAR|table is extended with the possibility of pressure sensing. A pair of load cells with working range up to 5 kg is placed below the front corners, as shown in Fig. 4.27. They are connected to an Arduino Pro Micro microcontroller via a pair of HX711 24 bit A/D-converters. Both sensors act as an electronic balance, powered via USB. The microcontroller is programmed using the Arduino language to provide a Universal Plug and Play (UPnP) USB MIDI device forwarding both load values as separate Control Change (CC) messages. The balances are set up and zeroed at startup, automatically adapting to the default load of the interface and possibly placed objects. The source code is available in Source 4.2.

</>

4.8. Discussion

With AltAR, we now have the technology to create plausible auditory augmentations for any rectangular flat surface. The most obvious use case is the augmented table. A demonstration video is provided in Source 4.3. The source code will be successively published in Source 4.4.

📄
</>

Practical applications in the form of sonifications based on AltAR/table (or parts of it) comprise the *Mondrian table* and the *auditory coloring book* in Sec. 8.1, as well as the prototypes for augmented writing (Sec. 7.3.1), exploratory data analysis (Sec. 7.3.2), and assisted object positioning (Sec. 7.3.7).

In the following chapter (Ch. 5), we will examine the perception of physical properties of rectangular plates: by robots listening to impact sounds (Sec. 5.1), by humans interacting with the AltAR/table (Sec. 5.2), and by humans listening to impact sounds (Sec. 5.3).

Bibliography

- Abel, Jonathan and David Berners (2004). “Filter Design Using Second-Order Peaking and Shelving Sections”. In.
- Alcan Composites (2006). *Technisches Datenblatt DIBOND*. ALCAN.
- Berdahl, Edgar and Dan Harris (2010). “Frequency Shifting for Acoustic Howling Suppression”. In: *International Conference on Digital Audio Effects (DAFx)*. Graz, Austria.
- Bovermann, Till, René Tünnermann, and Thomas Hermann (Apr. 2010). “Auditory Augmentation”. In: *International Journal of Ambient Computing and Intelligence* 2.2, pp. 27–41. DOI: 10.4018/jaci.2010040102.
- Farina, Angelo (2000). “Simultaneous measurement of impulse response and distortion with a swept-sine technique”. In: *AES Convention*. Paris, France: Audio Engineering Society.
- Lossius, Trond, Pascal Baltazar, and Théo de la Hogue (2011). “DBAP - Distance-Based Amplitude Panning”. In: *International Computer Music Conference (ICMC)*.
- Noisternig, Markus (2017). „Breitbandige Signalaufbereitung in Ein- und Mehrkanal-Mikrofonanwendungen“. Diss. Universität für Musik und darstellende Kunst Graz.
- Pirkle, Will C. (2019). *Designing Audio Effect Plugins in C++*. 2nd ed. Routledge. ISBN: 978-0-429-49024-8.
- Smith, Julius Orion (2010). *Physical audio signal processing: For virtual musical instruments and audio effects*. W3K publishing. ISBN: 978-0-9745607-2-4.
- Välämäki, Vesa and Joshua Reiss (May 6, 2016). “All About Audio Equalization: Solutions and Fron-

- tiers". In: *Applied Sciences* 6.5. DOI: 10.3390/app6050129.
- Vetter, Katja and Serafino di Rosario (2011). "ExpoChirpToolbox: a Pure Data implementation of ESS impulse response measurement". In: *Pure Data Convention*. Weimar, Germany.
- Waterschoot, Toon van and Marc Moonen (Feb. 2011). "Fifty Years of Acoustic Feedback Control: State of the Art and Future Challenges". In: *Proceedings of the IEEE* 99.2, pp. 288–327. DOI: 10.1109/JPROC.2010.2090998.
- Weger, Marian, Thomas Hermann, and Robert Höldrich (June 2018). "Plausible Auditory Augmentation of Physical Interaction". In: *International Conference on Auditory Display (ICAD)*. Houghton, Michigan, pp. 97–104. DOI: 10.21785/icad2018.024.
- (2022). "AltAR/table: a platform for plausible auditory augmentation." In: *International Conference on Auditory Display (ICAD)*. Virtual Conference.

5. Auditory perception and information capacity of rectangular plates

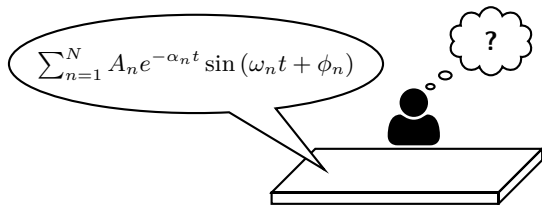


Figure 5.1.: How many bits fit in a rectangular plate?



Parts of this chapter are published in a more condensed form by Weger, Aurenhammer, Hermann, and Höldrich (2022).

In the previous two chapters, we created a novel auditory display that modulates the sonic appearance of a table in order to convey additional information. It uses a physical sound model to simulate the sound of a rectangular plate to simulate arbitrary physical parameters. How much information can be transmitted by this method? And more importantly, how much of the transmitted information are we able to perceive? This question is illustrated in Fig. 5.1. In other words, we want to know the information capacity of rectangular plates. We do not expect that the dimensions of a plate can be perceived as precise as the pitch of a sine tone. Nevertheless, we argue that for physical parameters, listeners can draw on their experience from everyday life, so that less training is required. Furthermore, people might even be able to absolutely identify different materials or shapes. In an auditory display on an otherwise continuous scale, such absolutely identified discrete levels might serve as anchor points for the user.

In order to answer our research questions, we will first examine how the physical properties of a rectangular plate are encoded in their sound, and how this physical information may be extracted by means of an algorithm for robotic auditory perception (Sec. 5.1). A two-dimensional auditory augmentation using length and aspect ratio, based on the AltAR/table platform is evaluated in Sec. 5.2. A follow-up listening experiment using pre-rendered sounds then investigates a three-dimensional display that employs aspect ratio, metallicity, and rigidity (Sec. 5.3). General conclusions on the informa-

tion capacity of plausible auditory augmentations of rectangular plates are drawn in Sec. 5.4.

5.1. An algorithm for robotic perception of material and dimensions of rectangular plates

This section is based on research performed by Czuka (2021) as part of his master thesis under my supervision, building on my original ideas. A summary article in German is provided by Czuka, Weger, and Höldrich (2021).



“Zillions have been spent to improve long-range underwater detection, but when some stumbling crew member in a submerged submarine (...) drops a wrench, only the human listener can identify the unexpected sound”

— Schroeder (1976, p. 184)

Great advances have been made in the robotic perception of material since the early attempts of Wildes and Richards (1988) and Krotkov (1995) (see Sec. 2.4 for a summary). However, since the high availability of machine learning algorithms such as deep neural networks, it seems that researchers of this new era sometimes forget the profound knowledge on acoustics that has been accumulated by generations of researchers until now. In Ch. 3 we derived a detailed physical model of plates and bars. We wonder what performance a detection algorithm could achieve if it could draw on all this knowledge.

In addition, not every combination of physical parameters of the model is actually physically feasible. For example, glass plates are extremely difficult to produce at larger thicknesses without cracking, due to thermal stress in the casting process. Such limits of physical feasibility can be applied to all model parameters. The previously unlimited parameter space is thus limited to a region that is physically feasible by means of strict laws of nature, but also the current state of human development, concerning the limits of state-of-the-art manufacturing processes.

This physically feasible parameter space can be further reduced by incorporating knowledge on the context such as environmental and cultural peculiarities. For example, in a densely-populated urban environment, wood and stone exist predominantly in cuboid shape, in form of planks or blocks. Contrary to their natural appearance, mainly in form of branched rods of wood in a forest, or raw lumps of rocks in the mountains. Similar to size or shape, also specific materials might dominate in a certain region or context while it is unlikely to find them in another. Different to the somehow hard limits of physical feasibility, this context-dependency is better described in terms of probabilities. It is absolutely possible to find a gold bar on the streets, but this is expected to happen with rather low probability.

We therefore propose an algorithm for robotic auditory perception of material and dimensions of rectangular plates, that exploits the knowledge of physical sound generation by incorporating the physical model described in Ch. 3 as well as feasibility- and context-dependent probabilities of certain combinations of physical parameters. This approach is intended to investigate what kinds of information on the physical object can be extracted from a single sound, resulting from an impact. In addition, by comparing the results with those of human listeners, we expect insights on human perception of physical properties.

In analogy to the implemented sound model, we assume that a single impact by a mallet produces a sound $s(t)$ that can be decomposed in the form

$$s(t) = \sum_{n=1}^N A_n e^{-\alpha_n t} \sin(\omega_n t + \phi_n) , \quad (5.1)$$

where n is an arbitrary mode number that denotes modes in ascending order with respect of their angular frequency $\omega_n = 2\pi f_n$.

A person who explores the physical properties of a plate by means of percussion, i.e., knocking on it in order to retrieve the relevant information from the resulting auditory feedback, would usually not always hit the same spot but rather explore it at different positions. This is actually important for correct identification, as there is no single spot on a plate where all modes are excited with equal weight. In case of identical boundary conditions on all four edges of a rectangular plate, each position shares its absolute amplitude distribution with three other equivalent positions on the plate, due to the two axes of symmetry. For free boundaries the border provides a good compromise for exciting all modes

roughly at equal amplitude. While there are still amplitude differences between modes, at least no mode exhibits a node at this position.

For robotic perception, we therefore assume that the plate is impacted at the corner, and also the microphone is placed nearby.

5.1.1. Measuring sound parameters: amplitudes, frequencies, decay factors

The first step in robotic auditory perception is the extraction of sound parameters from the recording. In principle, the natural frequencies f_n can be identified as the local maxima of the long-time average spectrum of the recorded impact sound. Note that the peaks actually occur at the damped frequency ω_r ; however, the difference to ω_0 is negligible in case of low damping. A single mode can then be isolated by a band-pass filter, under the assumption of low modal overlap. For each mode, a starting amplitude A_n and a decay factor α_n is fitted to the measured envelope to describe the assumed exponential decay.

5.1.1.1. Natural frequencies

For estimating natural frequencies, we use the fast partial tracking algorithm that was proposed by Neri and Depalle (2018). While there are many similar algorithms available, this one was chosen due to its availability¹, its robust performance, as well as its real-time capabilities which might facilitate a future real-time application. In comparison to our simple signal model from Eq. 5.1, the partial tracking algorithm uses a broader definition where the (instantaneous) amplitudes $a_n(t)$, frequencies $f_n(t)$, and phases $\phi_n(t)$ of partials are allowed to change over time:

$$s(t) = \sum_{n=1}^N \exp(a_n(t) + j\phi_n(t)) \quad (5.2)$$

with

$$\phi_n(t) = \phi_n(0) + 2\pi \int_{u=0}^t f_n(u) du . \quad (5.3)$$

These instantaneous signal parameters are computed for each frame k of the short-time Fourier transform (STFT) of the input signal.

¹Fast partial tracking in Matlab/Octave:
<https://github.com/jundsp/Fast-Partial-Tracking>

In order to allow arbitrary changes in frequency over time, including crossing partials, the detected partials of the individual signal frames are connected by applying a basic linear assignment problem (see Neri and Depalle 2018 for details). The algorithm achieves to track the majority of natural frequencies below 3 kHz. Above, the low-pass characteristic of the excitation as well as the internal and external damping make a robust detection impossible. Below about 50 Hz, the recorded sound is dominated by noise.

The most prominent partials detected by the tracking algorithm are picked for further processing. Assuming that the frequencies of partials do not significantly change over time, the instantaneous frequencies are averaged over all signal blocks, in order to obtain an average frequency f_n for each partial.

5.1.1.2. Modal weights and decay factors

The amplitudes and decay factors are estimated from the envelopes of the individual sinusoids. To extract these, the input signal is filtered by a Gaussian band-pass filterbank whose bands are tuned to the natural frequencies that have been obtained before.

The ideal Gaussian filter preserves the exponential decay if the input signal $s(t)$ is convolved with its impulse response $h_G(t)$. In general, the impulse response can be written as

$$h_G(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-t^2/2\sigma^2} \cos(2\pi f_c t) \quad (5.4)$$

with center frequency f_c and standard deviation

$$\sigma = \frac{\sqrt{-2\ln(G)}}{\pi B} \quad (5.5)$$

with filter cutoff gain G and filter bandwidth B .

We set the center frequencies f_c to f_n , while the bandwidths B_n are individually adjusted to the maximum possible value that allows sufficient separation between modes without too much ringing of the filters themselves:

$$B_n = 2 \min \{ |f_n - f_{n-1}|, |f_{n+1} - f_n| \} . \quad (5.6)$$

The cutoff gain was set to a constant value of $G = -40$ dB.

Each individual band-pass channel now allows the estimation of an envelope with the help of the Hilbert transform $\mathcal{H}\{\}$. The channel envelope is

described by

$$e_n(t) = \left| s_n(t) + j \mathcal{H}\{s_n(t)\} \right| . \quad (5.7)$$

A perfect exponential decay $e'_n(t)$, as assumed in our signal model,

$$e'_n(t) = A_n e^{-\alpha_n t} , \quad (5.8)$$

would lead to a linear envelope in the logarithmic domain:

$$\ln(e'_n(t)) = -\alpha_n t + \ln(A_n) . \quad (5.9)$$

The approximate slope and intercept of the envelope $e_n(t)$ are estimated by linear regression; their values readily provide estimations of the decay factors α_n and starting amplitudes A_n .

The acoustic measurements were verified on the basis of laser vibrometer measurements with exemplary plates of glass and aluminum. These measurements were in very high agreement (see Czuka 2021), so that we feel safe to say that the acoustic measurements can be fully trusted, especially if they are not used individually, but only statistically.

5.1.1.3. Critical frequency of radiation damping

In Sec. 3.2.6.1 we showed that radiation damping is the predominant form of damping at high frequencies above a certain critical frequency f_{cr} . For isotropic plates it can be derived from physical properties of the plate (see Eq. 3.74), and might thus serve as an additional equation for our under-determined equation system in the task of inferring from sound properties to physical properties.

If all other forms of damping are negligibly low in comparison to radiation damping, we should be able to derive the critical frequency based on measured decay factors. Unfortunately, the measurement of decay factors of individual partials was only possible up to a certain frequency that might be lower than the critical frequency. We therefore compute average decay factors measured in frequency bands to obtain an estimation of the frequency-dependent damping due to radiation. As we are only interested in the critical frequency, and not in the actual amount of damping, an under- or overestimation of the average decay factors will not change the result, as long as the error is approximately constant across frequency bands. We employ a third-octave filterbank according to IEC 61260-1:2014 (International Electrotechnical Commission 2014). The result is a

5. Auditory perception and information capacity of rectangular plates

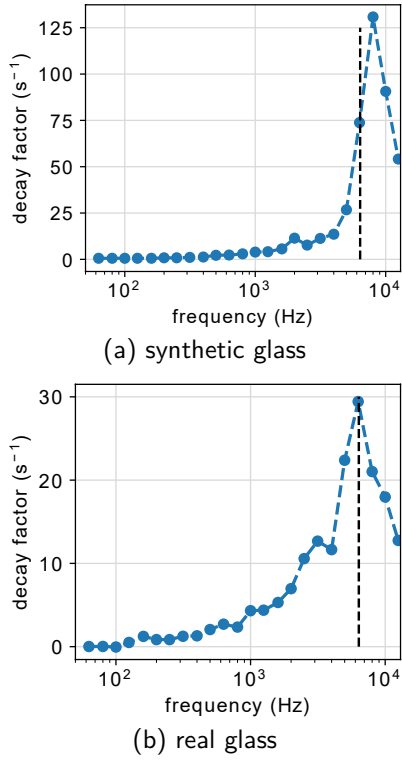


Figure 5.2.: Measured decay factor in 1/3-octave bands and estimated critical frequency, for a synthesized and real glass plate. Illustration by Czuka (2021).

measured decay factor for every third-octave band. Based on synthesized sounds, the center frequency of the band below the global maximum provided a sufficient approximation for the critical frequency f_{cr} of the radiation damping (see Fig. 5.2).

5.1.2. Estimating physical parameters from sound parameters

Based on the measured sound parameters in Sec. 5.1.1, we are now able to deduce the underlying physical properties of the impacted rectangular plate.

5.1.2.1. Internal damping and material

Below the critical frequency f_{cr} , the damping due to radiation is relatively low, in comparison to the internal damping mechanisms (viscoelastic, viscous, and thermoelastic). The measured decay factors below the critical frequency are therefore assumed to provide a robust predictor of internal damping.

Based on the damping model in Sec. 3.2.6, we assume the decay factors to follow the line equation

$$\alpha(f) = \eta_v \pi f + \alpha_f, \quad (5.10)$$

where the slope $\eta_v \pi$ represents the viscoelastic (frequency-dependent) part of the damping, whereas the intercept α_f may include both thermoelastic and viscous damping (see also the relationships between damping parameters in Sec. 3.1.4).

A linear regression on the measured damping factors directly provides estimates for η_v and α_f . Figure 5.3 shows exemplary measurements for a synthesized and a real aluminum and glass plate, respectively. The regression lets us even conjecture about the metallicity of the plate, i.e., the amount of thermoelastic damping. Pure viscoelastic damping follows a straight line, which leads to a strong correlation between the simple line model of decay factors and the measurement. Thermoelastic damping varies greatly between individual modes, especially for the lower ones. The coefficient of determination R^2 may therefore serve as a descriptor for metallicity, and thus allow a discrimination between materials of equal Young's modulus and density, such as glass and aluminum.²

5.1.2.2. Dimensions and boundary conditions

From Sec. 3.2.5 we know that the (undamped) natural frequencies can be defined in terms of a base frequency Φ and non-dimensional frequency factors λ_n that describe the frequencies of the partials relative to the base frequencies. Let us recapitulate these briefly. According to Eq. 3.59, the natural frequencies are

$$f_0 = \frac{\omega_0}{2\pi} = \frac{\Phi \lambda}{2\pi} \quad (5.11)$$

with

$$\Phi = \frac{\pi h}{2S} \sqrt{\frac{D}{\rho}} = \frac{\pi}{\sqrt{48}} \frac{hc_L}{S} \quad (5.12)$$

and

$$\lambda = \left[\frac{1}{r_a^2} G_x^4 + 2\nu H_x H_y + r_a^2 G_y^4 + 2(1-\nu) J_x J_y \right]^{1/2}. \quad (5.13)$$

²Remember the damping interpolation parameter H in Sec. 3.2.6.5.

5.1. An algorithm for robotic perception of material and dimensions of rectangular plates

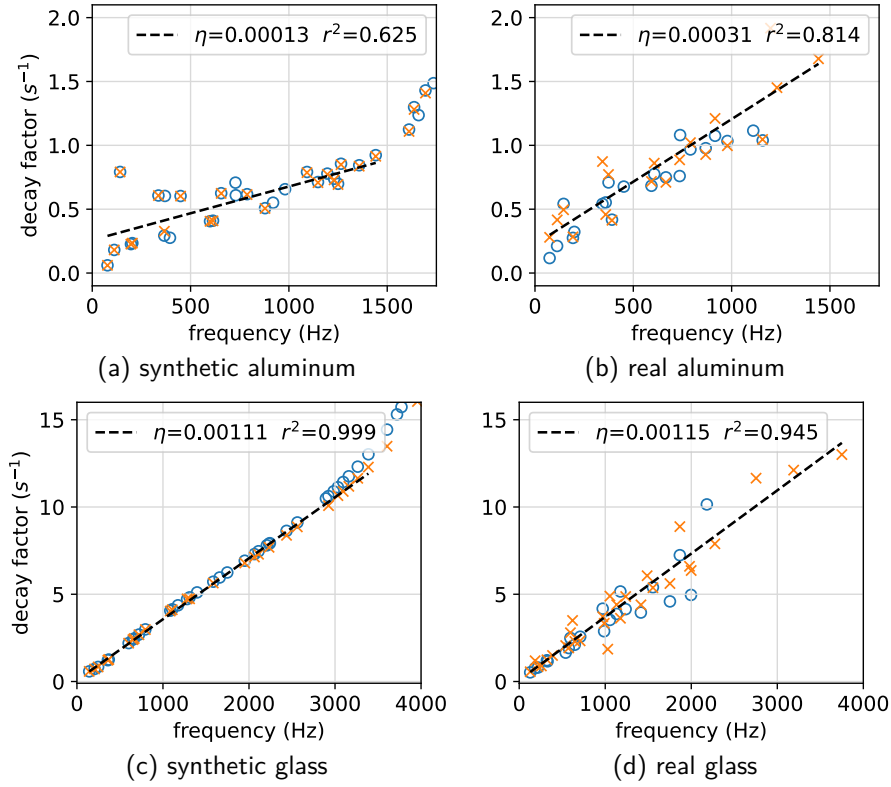


Figure 5.3.: Measured decay factors α and estimated loss factor η of an impacted plate. Illustration by Czuka (2021).

Φ depends on size (thickness h and area S) and material (density ρ and rigidity D which can be further fractionized into Young's modulus E and Poisson's ratio ν). λ depends on the aspect ratio r_a , the boundary conditions (via G , H , and J), the material (Poisson's ratio ν), and the individual mode indices m and n . In practice, ν has no significant effect on the frequency ratios and can simply be set to an average value of ≈ 0.35 that is common to "most materials" (Rossing 2014, p. 590).

For a certain combination of boundary conditions, the measured frequencies must therefore match a pattern that is defined only by Φ and r_a . In other words, the one combination of Φ and r_a that leads to frequencies that fit best to the measured frequencies might be a good estimate of the true value of Φ and r_a .

It is obvious that the number of measured frequencies \tilde{f}_i is smaller than the (infinite) number of model frequencies f_j . In fact, we do not know which modes are missing, and we might lack not only some lowest modes (not radiated) and some higher modes (strongly damped), but also some modes in between which are not excited or just not

captured due to destructive interferences. Therefore, every measured frequency \tilde{f}_i is exclusively assigned to a model frequency f'_j by solving a linear assignment problem by means of the so-called Hungarian algorithm (Kuhn 1955).

Each of the I measured frequencies is assigned exclusively to one of the J model frequencies. The assignment problem can be written as

$$\min \left\{ \sum_{i=1}^I \sum_{j=1}^J C_{ij} \right\} \quad (5.14)$$

with

$$\sum_{i=1}^I X_{ij} = \sum_{j=1}^J X_{ij} = 1, \quad (5.15)$$

where the binary values for assigned (1) and not assigned (0) are stored in matrix \mathbf{X} .

The assignment problem is solved by minimizing the cost function, with the elements C_{ij} of the cost matrix \mathbf{C} defined as

$$C_{ij} = \frac{|\tilde{f}_i - f'_j[k]|}{\tilde{f}_i}. \quad (5.16)$$

5. Auditory perception and information capacity of rectangular plates

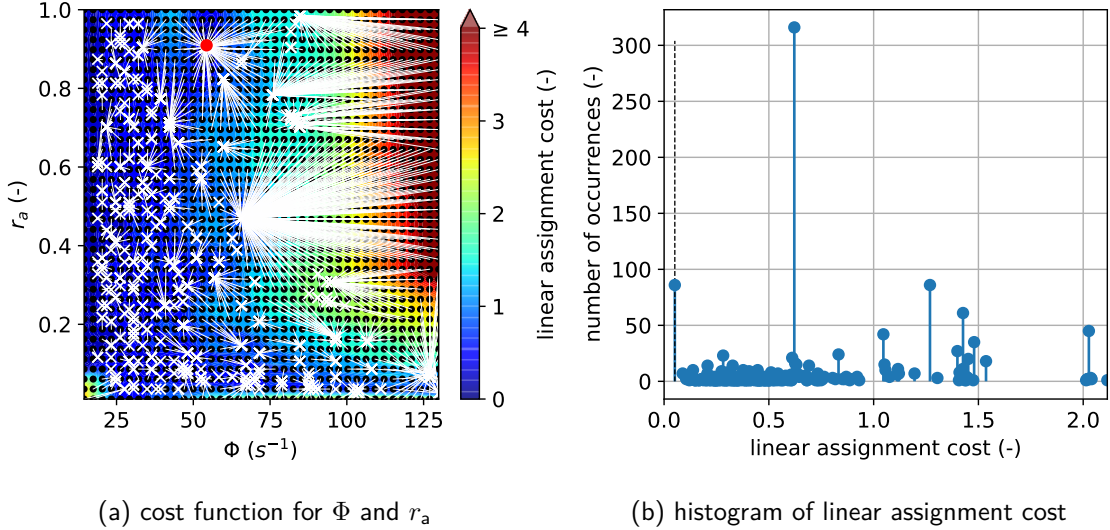


Figure 5.4.: (a) Cost function with true values (red dot) and paths of the gradient search (white lines). (b) Histogram of the linear assignment cost (dashed line marks best fit). Illustration by Czuka (2021).

A steepest descent algorithm is applied to find the best-fitting model parameters Φ' and r'_a so that the error between model frequencies f'_j and measured frequencies \tilde{f}_i is minimized.

For each pair of initial values, the assignment which induces the least assignment cost is obtained. For the next iteration of the steepest descent algorithm, a new pair of values $\Phi'[k+1]$ and $r'_a[k+1]$ is derived according to

$$\Phi'[k+1] = \Phi'[k] - \mu_{\Phi} \frac{\partial \min \left\{ \sum^I \sum^J C_{ij} X_{ij} \right\}}{\partial \Phi} \quad (5.17)$$

and

$$r'_a[k+1] = r'_a[k] - \mu_{r_a} \frac{\partial \min \left\{ \sum^I \sum^J C_{ij} X_{ij} \right\}}{\partial r_a}, \quad (5.18)$$

with the associated learning rates μ_{Φ} and μ_{r_a} . The algorithm leads to an error surface which depicts the minimum of the cost function of the linear assignment problem as a function of Φ' and r'_a .

Starting from a grid of meaningful initial values $\Phi'[0]$ and $r'_a[0]$ (e.g., with $\Phi'[0]$ between 25 and 225 Hz, and $r'_a[0]$ between 0.1 and 1.0), the algorithm iterates along both axes of the error function until a local minimum of the cost function is reached. Each of the local minima that have been found this

way, is then characterized in terms of its total linear assignment cost and the number of starting points that led to it, i.e., its hit frequency.

Results are shown in Fig. 5.4 for a synthesized aluminum plate. Obviously, the synthesized plate is based on the same model as the fitted model, so that the best fit approaches the measured frequencies with very low linear assignment cost. Nevertheless, the algorithm is still capable of finding the true values Φ and r_a even for recordings of physical plates; however, at a larger assignment cost. This larger cost is attributed mainly to the fact that the mode shapes of the real plates diverge from those in the model which are based on the approximate solutions by Warburton (1954).

Note that a low value of Φ leads to a high modal density $DM(f)$, with the effect that for any measured frequency, there will be some model frequency nearby to assign to. In the extreme, with Φ approaching zero, the modal density tends towards infinity, so that also the assignment cost becomes zero for any aspect ratio r'_a . Therefore, an estimate of the modal density is computed, leading to a lower limit of the starting values $\Phi'[0]$. This estimate is simply the inverse of the average frequency distance between neighboring natural frequencies, or the frequency range divided by the number of modes:

$$DM(f) = \frac{1}{N} \sum_{n=1}^{N-1} (f_{n+1} - f_n) = \frac{f_N - f_1}{N - 1}. \quad (5.19)$$

Table 5.1.: True and estimated loss factors η_v .

ID	type	material	true	estimated	R^2
			$\eta_v (\times 10^{-4})$	$\eta_v (\times 10^{-4})$	
a_1	synth.	aluminum	[2.2, 7.7]	2.1	0.33
a_1	real	aluminum	[2.2, 7.7]	2.8	0.43
a_3	synth.	aluminum	[2.2, 7.7]	1.3	0.63
a_3	real	aluminum	[2.2, 7.7]	3.1	0.81
v_1	synth.	glass	[6, 20]	11	1.00
v_1	real	glass	[6, 20]	12	0.95
c_1	synth.	carbon	—	21	0.33
b_1	synth.	spruce	≈ 80	76	0.64

In order to obtain an estimate of the boundary conditions, the assignment problem can be solved for all 15 possible combinations of basic boundary conditions. The best-fitting combination is then selected.

Now that we have an estimate of frequency factor Φ and f_{cr} , we can reformulate both corresponding equations based on only three unknowns: longitudinal wave velocity c_L , thickness h , and surface area S . Similar to Φ in Eq. 5.13, the critical frequency can be rewritten as

$$f_{cr} = \frac{1}{2\pi} \frac{c_0^2}{h} \sqrt{\frac{\rho}{D}} = \frac{\sqrt{3}c_0^2}{\pi} \frac{1}{hc_L}. \quad (5.20)$$

Both can be reformulated with respect to hc_L . Inserting one into the other yields an estimation for the surface area:

$$S = \frac{c_0^2}{4f_{cr}\Phi}. \quad (5.21)$$

From aspect ratio and surface area, we can now compute the exact length and width of the plate:

$$l_x = \sqrt{S \cdot r_a}, \quad l_y = \sqrt{S/r_a}. \quad (5.22)$$

Unfortunately, it is not possible to separate the thickness h from the longitudinal wave velocity c_L . At least, however, their product can now be estimated via Eq. 5.13:

$$hc_L = \frac{\sqrt{48}}{\pi} \Phi S. \quad (5.23)$$

The measured frequencies unfortunately do not contain any useful information about the thickness h and longitudinal wave velocity c_L , with both being inseparably integrated in Φ and f_{cr} .

An analysis of plausible parameter ranges, however, might give a hint for their separation. For each

material category and shape (width and length), there is a plausible range of thicknesses that are technically feasible and also likely to appear (a gold plate might be feasible but unlikely in most contexts).

5.1.3. Case studies

The algorithm for estimating physical parameters from sound recordings of impacted plates was tested on a number of synthetic plates as well as on recordings of real physical plates. The real plates are replica of the aluminum plates a_1 and a_3 , and the glass plate v_1 used by Chaigne and Lambourg (2001). The synthesized plates include a_1 , a_3 , and v_1 , as well as carbon fiber plate c_1 and wooden plate b_1 (spruce), with the physical parameters (material and dimensions) given by Chaigne and Lambourg. With plates c_1 and b_1 being orthotropic, we assume that the algorithm (in its current form covering only isotropic plates) will run into problems. The true elastic constants of the individual plates were not measured but instead taken from the literature for the given material.

The estimated loss factors are summarized in Tab. 5.1. The true loss factors are based on parameter ranges given by Cremer et al. (2005, pp. 191–196). The measured loss factors of the real plates were always within the given range for the specific material (aluminum, glass, and spruce³). A rough guess of the materials may thus be possible on the basis of the loss factor alone — if the inspected plate is known to be undamped, e.g., freely suspended as in our case for free boundary conditions. In case of additional damping (e.g., by holding the plate in the hand), the true damping values given in the textbooks don't apply anymore. The coefficient of

³taking the value for fir given by Cremer et al. (2005, p. 196)

5. Auditory perception and information capacity of rectangular plates

Table 5.2.: True and estimated values for critical frequency f_{cr} , frequency factor Φ , aspect ratio r_a , and area S , based on measurements on synthesized and real physical plates.

ID	type	material	true				estimated			
			f_{cr}/kHz	r_a	Φ/Hz	S/m^2	f_{cr}/kHz	r_a	Φ/Hz	S/m^2
a_1	synth.	aluminum	6.00	0.63	83.5	0.059	6.30	0.62	79.3	0.059
a_1	real	aluminum	—	0.63	—	0.059	6.30	0.76	78.48	0.059
a_3	synth.	aluminum	3.05	0.91	54.4	0.178	3.15	0.90	55.1	0.170
a_3	real	aluminum	—	0.91	—	0.178	3.15	—	54.5	0.171
v_1	synth.	glass	6.37	0.97	89.7	0.052	6.30	0.93	96.5	0.048
v_1	real	glass	—	0.96	—	0.052	5.00	—	—	—
c_1	synth.	carbon	6.37	0.50	57.1	0.080	8.00	—	—	—
b_1	synth.	spruce	5.04	0.8	27.5	0.212	5.00	—	—	—

determination R^2 gives another hint on the material. Taking only isotropic ones into account (aluminum and glass), a hard threshold at about 0.9 might already lead to a robust discrimination between viscoelastic and thermoelastic damping; however, this needs to be verified on the basis of a larger dataset.

The synthesized carbon fiber and spruce plates lead to relatively low values of R^2 , similar to the aluminum plates. However, in this case, the deviation from the straight line for the decay factors α is not due to thermoelastic damping, but due to orthotropy. A high coefficient of determination can therefore only come from an isotropic and viscoelastic material, but a low coefficient of determination can have several reasons (e.g., thermoelastic damping, orthotropy, or just selective damping of some individual modes due to an external damping mechanism).

The results of the frequency matching algorithm for estimating frequency factor Φ and aspect ratio r_a are given in Tab. 5.2 together with estimates of critical frequency f_{cr} and the derived area S of the plate. First of all, it can be observed that for the orthotropic materials, the steepest descent algorithm failed in finding an optimal solution to the assignment problem. Hence, Φ and r_a could not be estimated for these plates. While these plates are not covered by the model behind our algorithm, it failed also completely for the real recording of the glass plate v_1 , and partially (for Φ) for the real aluminum plate a_3 . In all other cases, however, the found solution is very close to the true value.

5.1.4. Discussion

For isotropic rectangular plates, we were able to recover size and shape (aspect ratio, area, length,

width) with very high precision. As the proposed algorithm is based on a model of isotropic plates, it did not converge for orthotropic plates.

In case that the examined plate is vibrating freely, i.e., without any external damping (except radiation damping), as is approximately the case for a plate that is hanging on thin strings, the measured loss factor directly leads to the material category via tables of measured loss factors as given by Cremer et al. (2005, pp. 191, 195–196). However, this is a rather unrealistic and impractical assumption. In practice, i.e., for more complex boundary conditions, inducing additional external damping, the measured loss factors will be much larger than those ideal ones given in the literature, so that these cannot contribute to a reliable material estimate. Additional external damping can only add to the damping of freely vibrating plates. This means that a loss factor of 15 can be undamped glass, but also aluminum with extra damping. On the other hand, it cannot be spruce, as this has already a value of 80 in the undamped case. In practice, the larger the measured loss factor, the larger the possible range of materials. A small loss factor limits the number of possible materials.

In the current implementation, the algorithm does not take into account the deformation of mode shapes in case of modes with the same natural frequency, occurring at aspect ratios of 1/1, 1/2, 1/3, etc. In this case the fitting of the model frequencies is unreliable. The same applies for extremely small aspect ratios ($r_a \ll 1$, or extremely large with $r_a \gg 1$), where the plate effectively becomes a bar. Due to the lower modal density of bars, the fitting algorithm leads to more ambiguous results for aspect ratio and area.

McIntyre and Woodhouse (1988) derived approx-

imation formulas to estimate Poisson's ratio and Young's modulus from the frequencies of the ring- (o) and \times -modes which only appear at aspect ratios close to 1:

$$\nu \approx 1.48 \frac{f_o^2 - f_x^2}{f_o^2 + f_x^2} \quad (5.24)$$

and

$$E \approx 0.46 (1 - \nu^2) (f_o^2 + f_x^2) \frac{\rho S^2}{h^2} . \quad (5.25)$$

Nevertheless, even in this special case, due to the unknowns ρ and h in Eq. 5.25, E cannot be directly computed. A reformulation again yields the inseparable pair hc_L that can only be derived together:

$$(hc_L)^2 = h^2 \frac{E}{\rho(1 - \nu^2)} \approx 0.46 (f_o^2 + f_x^2) S^2 . \quad (5.26)$$

If an aspect ratio of 1 is measured, however, Eq. 5.24 is a convenient method to obtain an estimation of Poisson's ratio ν and thus an additional hint on the true material.

5.2. Multisensory discrimination of size and aspect ratio



This experiment is connected to the bachelor thesis by Aurenhammer (2021) who implemented the pseudo-random trial order under my supervision and helped conducting the experiment.

Our personal experiences suggest that the average person is able to distinguish between different materials and shapes of rigid physical objects by exploring the auditory feedback through tapping, or scratching. These sonic skills are surely not even limited to a certain culture such as our western civilization, but can be attributed to more or less all human beings on the planet. On the basis of listening experiments with synthesized and also with physically struck objects (see Sec. 2.2), we feel safe to say that humans can almost perfectly discriminate between gross material categories (glass/metal vs. wood/-plastic, at least for small damping), and to some extent even between materials within categories (e.g., glass vs. metal) or between different sizes and shapes (e.g., small vs. large, or plate vs. bar). In addition, we know that the perception and discrimination of physical parameters benefits from the

combination of different sensory modalities such as audition, vision, and touch.

Based on general knowledge in acoustics, namely the physical model presented in Ch. 3, we are able to infer the connection between physical properties and sound parameters. In Sec. 5.1 we showed to what extent it is possible to reverse this process, with the aim to derive physical parameters from measured sound parameters. A major problem is the fact that this direction defines an under-determined problem, with an unlimited range of possible physical parameters for a certain combination of distinct sound parameters. For example, size and material both affect pitch, which leads to a confound between both parameters. Listening experiments have shown that we tend to base our judgments on expectations due to our everyday acoustic environment (e.g., small metal bars and large glass plates, see Sec. 2.2).

In addition, even if a certain sound parameter is connected to only one single physical parameter, our perceptual resolution differs across sound parameters, hence leading to different perceptual resolution across physical parameters (see Sec. 2.3).

Incorporating all these aspects, we designed an experiment to explore the perception of size and shape of rectangular plates in an ecological scenario of percussion. We want to investigate to what extent participants are able to distinguish between size (here in the form of length) and shape (here in the form of aspect ratio), and additionally, with what precision participants are able to estimate the size and shape of rectangular plates of different materials. We assume that most people in our western society are familiar with the sounds of rectangular plates, based on their experiences from everyday life. In a building material store, almost everything is rectangular. In our homes and in our offices, many objects are rectangular or constructed from rectangular parts (even this thesis is rectangular). It might be possible that we are experts in auditory rectangularity. For sure, we expect individual differences. As an example, the beggar in front of a bank might be more experienced in auditory discrimination of coins than the banker inside. However, as we are limited to a small number of participants, we do not expect to cover such social or cultural aspects experimentally.

The technical setup is based on the AltAR/table platform which is described in detail in Ch. 4. During the experiment, participants directly interacted with the interface plate and identified an unknown plate's length and aspect ratio in direct comparison to a reference plate with known dimensions. The

5. Auditory perception and information capacity of rectangular plates

Table 5.3.: Model coefficients of the rendered plates.

		glass	wood
thickness	h / mm	10	12
viscoelastic loss	η_v or η_{iv}	0.001	[0.0051 0 0.0216 0.0164]
density	ρ / kg m ⁻³	2550	415
rigidities	D_i / MPa	[6700 - - 10 270]	[1320 77 82 227]
viscous loss	α_f / Hz	1.76 + 4	4.8 + 1
hardness	HB / kgf	1550	1.3

Table 5.4.: Levels of the independent parameters length and aspect ratio. Main levels are labeled by their short names, all others are distractor levels.

length / m		aspect ratio	
	0.297	1.100	
small	0.343	1.416	compact
	0.396	1.822	
medium	0.456	2.345	longish
	0.526	3.018	
large	0.607	3.884	bar-shaped
	0.700	5.000	

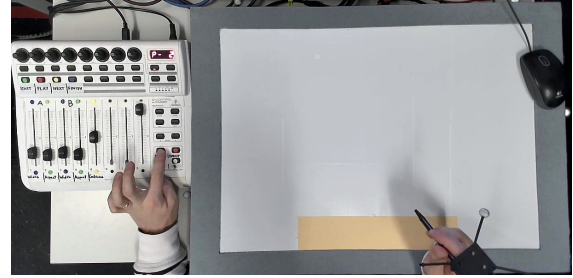


Figure 5.5.: Apparatus for experiment 1, including the interface plate, tracked pencil, and MIDI controller (captured via webcam during the experiment).

experiment was performed separately for the two materials glass and wood.

5.2.1. Stimuli

The model parameters of the rendered plates are given in Tab. 5.3. They were kept constant throughout the experiment while the length, i.e., the longest physical dimension of the plate, and the aspect ratio, i.e., relative width, were varied. Length and aspect ratio took 7 levels each, evenly spaced on a logarithmic scale. Their values are given in Tab. 5.4. Four of these are distractor levels; their only use was to lead the participants into believing that the parameters could take any value within the parameter range. The remaining three, we refer to them as main levels, are those that were actually tested. The two outer distractor levels define the minimum and maximum of the parameter range, respectively.

Both length and aspect ratio were jittered randomly for each new unknown plate to pretend an interval scale and to prevent participants from simply remembering distinct sounds, as is common practice in such listening experiments (e.g., Lutfi and Liu 2007). While main levels were jittered uniformly within $\pm 30\%$ of the logarithmic increment between adjacent levels, distractor levels were jittered within $\pm 40\%$, with the exception of the outmost distractor

levels which were only jittered inwards.

The material-inherent viscous damping α_f was extended by an additional value to equalize the overall decay times between materials to a similar value, as well as to shift the unfamiliar low damping of freely vibrating plates to a more ecologically valid range. The perceived damping of the virtual plates thus approximately equaled the physical plates that were used for the familiarization task (see Sec. 5.2.3).

The radiation efficiency was applied in a simplified form to model a more near-field behavior which was assumed to be more ecologically valid in the given situation. The fourth root of the radiation efficiency seemed a good compromise to accomplish this goal.

For wood, the model parameters describe an orthotropic material, with $\Omega = (D_1/D_3)^{1/4} = 2$ and a fiber direction parallel to the long edge of the plate.

5.2.2. Apparatus

Participants were seated in front of the table and interacted with the interface plate through a ballpoint pen that was equipped with infrared markers for optical tracking (see Fig. 5.5). Interaction was only possible within a small region of 297 mm \times 59.4 mm which equaled the dimensions of the smallest plate

that was modeled during the experiment. Within this region, a paper overlay of the same dimensions was placed. The excitation signal was only fed to the physical model if the tip of the pen was located within the active region.

The experiment itself was implemented in Pd which was running on a separate computer and sent control data and model parameters to the auditory augmentation system. The experiment GUI was shown on a screen; however, participants navigated through the experiment only by using the buttons of a Behringer BCF2000 MIDI controller. The motorized faders copied the GUI sliders for length and aspect ratio of the model plate and the unknown plate, as explained further in Sec. 5.2.3.

For every trial, audio and video recordings were made. These include all audio input and output channels within SuperCollider as well as a measurement microphone from above the interface plate. Interactions with the interface plate were captured by a webcam placed above; the recording itself was done in OBS Studio⁴ which was remote-controlled by the experiment software through keyboard shortcuts that were simulated via AutoHotkey⁵. Position tracking data was recorded in SuperCollider.

The experiment took place in an acoustically treated room of 4.3 m × 6.2 m × 3.4 m (length × width × height) size.

5.2.3. Procedure

The experiment was structured into four parts. First, participants were seated at a separate desk for instructions and familiarization. Within the written instructions, they were asked to tap on two prepared physical plates of glass and wood by using the same tracked ballpoint pen that was used later on the actual interface. Participants had to hit the plates at 3 distinct positions that were marked on a paper overlay in order to raise their awareness towards the influence of excitation position on the resulting sound.

For the next parts, participants took a seat at the augmented table. During passive training, the effect of length and aspect ratio was demonstrated by playing pre-recorded excitations through the interface. Both parameters stepped through the 7 levels, i.e., the whole range of the slider, from low to high and back, while the other parameter was set to a constant medium value, respectively. They

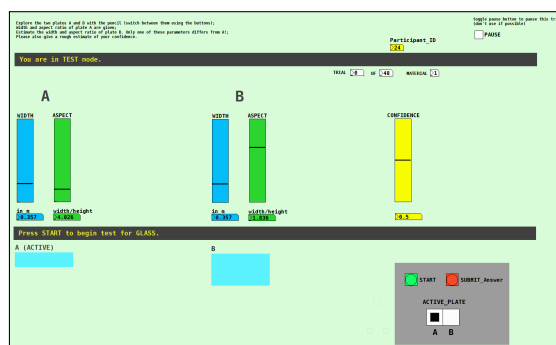


Figure 5.6.: Experiment software on the screen, in test mode, for experiment 1.

had to listen to all 4 combinations of parameter and material at least once to be able to proceed to the active training.

Active training and the actual test shared the exact same procedure, but with fewer trials and more salient parameter changes. Figure 5.6 shows the experiment software that was displayed to the participants on a computer screen. The training/test procedure was performed separately for both materials (glass and wood); with balanced order across participants. At the start of each trial, the augmentation was set to the reference plate A whose length and aspect ratio were given by two sliders (and motor faders). Participants could switch to the unknown plate B through a button. The state of length and aspect ratio was copied from the reference plate at start. Participants were asked to identify the parameter that differed between A and B, and set the corresponding slider to an estimated value. If one slider was moved, the other was instantly reset to the value of the reference plate to ensure an answer in only one parameter. Participants could change between both plates at will before submitting their answer and proceeding to the next trial. If participants decided for the wrong parameter, the background color switched to red, and participants were asked to correct their judgment. After responding in the correct parameter dimension, the background switched back to green and the next trial was presented.

To speed up the whole process, the unknown plate B of a trial stayed as reference plate A in the next trial. This included the resolution of the previous trial as feedback, and let participants directly proceed to the unknown plate B as they were already familiar with A.

The test of each material was therefore organized

⁴OBS Studio: <https://obsproject.com/>

⁵AutoHotkey: <https://www.autohotkey.com/>

5. Auditory perception and information capacity of rectangular plates

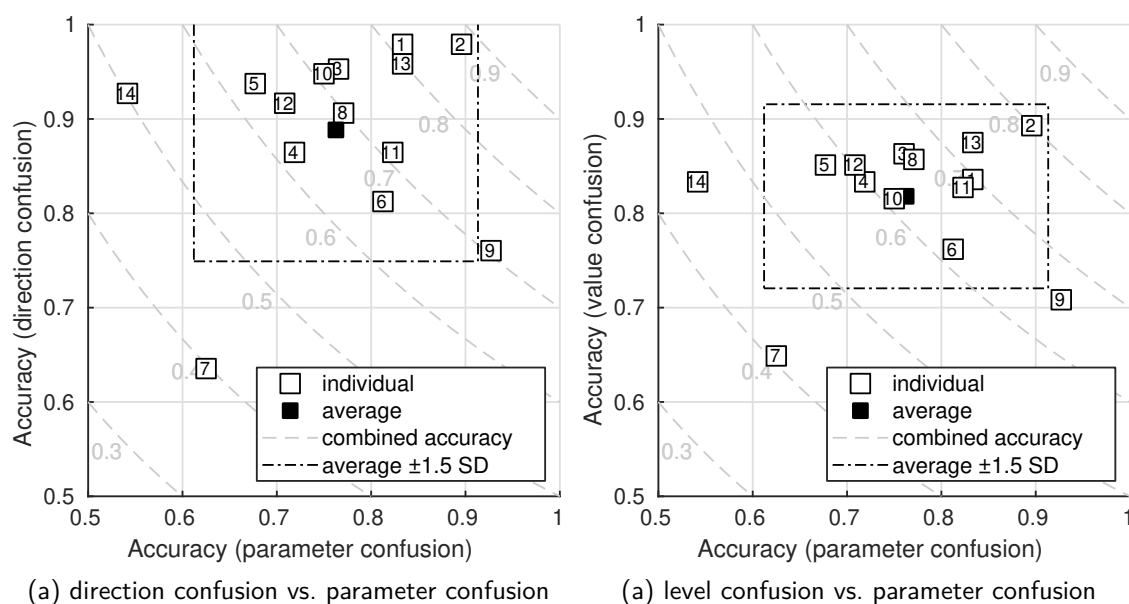


Figure 5.7.: The individual participants' performance: accuracy in parameter identification (length and aspect ratio), accuracy in level identification (3 parameter levels), and accuracy in direction identification (increase and decrease).

in a series of 48 trials which formed a trajectory through the 2D parameter space where only one parameter changed between successive trials (i.e., between A and B). The trajectories were pre-computed in Matlab so that each of the 9 combinations of main levels was reached as the unknown plate by all 4 combinations of main levels that were possible for the corresponding reference plate. This led to a total of 36 main trials per material.

In addition, distractor trials were generated which appeared always in pairs so that a parameter changed to a random distractor level and then returned back to a random main level. Six such pairs of distractor trials were inserted at random positions within the trajectory, but exactly once in a row of 9 main trials. There was always at least one main trial between two pairs of distractor trials.

5.2.4. Participants

A total of 14 participants (8 female, 6 male) were recruited to form a diverse mix of experts (4 colleagues, 4 graduate students in sound design) and non-experts (4 undergraduate students, family members, and friends). They received no compensation for their participation; all reported normal hearing.

5.2.5. Results

The collected data (responses, meta data, and tracking data) were imported to Matlab which was used for the statistical analysis. A general level of 5% was used as threshold for statistical significance.

5.2.5.1. Outliers and individual participants' accuracy

From the participants' individual confusion matrices on parameter dimension (length or aspect ratio), parameter value (3 levels per parameter), and parameter direction (increase or decrease), accuracies have been derived. For sonification, parameter confusion and direction confusion are assumed to be the most critical. In Fig. 5.7a, these are plotted against each other. Figure 5.7b shows the same plot with level confusion which might be critical if absolute identification is more important than the change over time, depending on the sonified data.

Average accuracies are 0.888 ($SD = 0.093$) for direction confusion, 0.818 ($SD = 0.065$) for level confusion, and 0.763 ($SD = 0.100$) for parameter confusion. The participants P7, P9, and P14 are more than 1.5 standard deviations away from the average in at least one of the three mentioned categories of accuracy. They are considered as outliers who will be removed from further statistical analysis

Table 5.5.: Confusion matrix for the parameter dimension of a change, for both materials separately, pooled over participants. Distractor plates are included.

(a) glass

		true	
		length	aspect
selected	length	184	59
	aspect	72	213

$Acc=0.75, \chi^2(1)=133.7$

(b) wood

		true	
		length	aspect
selected	length	209	32
	aspect	69	218

$Acc=0.81, \chi^2(1)=206.4$

that is based on pooled data.

5.2.5.2. Confusion between parameter dimensions

A fundamental task for the participants was to decide which of the two parameter dimensions (length or aspect ratio) had changed between A and B. Each trial can be attributed to one of 4 fields in the confusion matrix of true parameter and selected parameter. Table 5.5 shows these confusion matrices for parameter confusion, pooled over participants, for both materials. Overall accuracies (i.e., probabilities for choosing the correct parameter) were 0.75 for glass and 0.81 for wood.

5.2.5.3. Discrimination between main levels of length and aspect ratio

For comparison between main levels of length and aspect ratio, we exclude those trials in which the unknown plate B involves a distractor. Trials with a distractor as reference plate A, however, are included to maximize the sample size. Only the last answer of each trial is taken into account (the either correct or corrected one). The estimated values are jitter-corrected and rounded to the nearest main level. On the basis of these estimated main levels, confusion matrices are constructed for each parameter and material, see Tab. 5.6 and 5.7. The overall accuracy for length in case of wood is a bit higher ($Acc=0.87$)

Table 5.6.: Confusion matrix for the 3 main levels of length, for both materials separately, pooled over participants. Answers are jitter-corrected and rounded to the nearest main level on a logarithmic scale. Distractor plates are excluded, but the reference may be a distractor.

(a) glass

		true		
		small	medium	large
estimated	small	130	7	7
	medium	10	118	13
	large	6	33	138

$Acc=0.84, \chi^2(4)=539.1$

(b) wood

		true		
		small	medium	large
estimated	small	147	23	1
	medium	5	115	13
	large	1	18	139

$Acc=0.87, \chi^2(4)=603.0$

Table 5.7.: Confusion matrix for the 3 main levels of aspect ratio, for both materials separately, pooled over participants. Answers are jitter-corrected and rounded to the nearest main level on a logarithmic scale. Distractor plates are excluded, but the reference may be a distractor.

(a) glass

		true		
		compact	longish	bar
estimated	compact	123	13	11
	longish	19	129	11
	bar	4	18	134

$Acc=0.84, \chi^2(4)=528.3$

(b) wood

		true		
		compact	longish	bar
estimated	compact	131	17	4
	longish	19	127	24
	bar	3	12	125

$Acc=0.83, \chi^2(4)=522.7$

than for all other combinations (between 0.83 and 0.84).

Pooled over participants, the distributions of es-

5. Auditory perception and information capacity of rectangular plates

timated values for the main levels of length and aspect ratio, are unlikely to follow a normal distribution. The Lilliefors test was able to reject the null hypothesis of normally distributed data for most combinations of parameter, main level, and material. For the average levels of length and aspect ratio, i.e., medium and longish, respectively, we even expect at least some kind of bimodal distribution. We generally assume dependent data, as each combination of length, aspect ratio, and material was tested by the same participants. However, due to unequal sample sizes between some combinations, we take the pill of lower statistical power that comes with the attribution to unpaired data and stick to Mann-Whitney U tests for pairwise comparisons between median values. If not stated differently, any given p -values are Bonferroni-Holm adjusted.

Pooled over participants and lengths, a larger value of true aspect ratio always led to a significantly larger estimated aspect ratio for both materials. In particular, medium and large were perceived significantly larger than small, and large was perceived significantly larger than medium (all $p < 0.001$, respectively, for both materials).

For aspect ratio, the results follow the same trend. In particular, medium and large were perceived significantly larger than small ($p < 0.001$, respectively, for both materials), and large was perceived significantly larger than medium ($p = 0.026$ for glass, $p < 0.001$ for wood).

5.2.5.4. Direction and amount of a parameter change

Confusion matrices for sign of change of a parameter are constructed based on all trials, including distractors. They are shown separately for both parameters and materials in Tab. 5.8. While overall accuracies are generally higher, a similar trend as before before can be observed. In particular, accuracy for length in case of wood (0.97) was significantly higher than for the other combinations (0.89 and 0.92).

5.2.5.5. Estimated vs. true values

Figure 5.8 shows the average estimated lengths and aspect ratios for both materials separately, pooled over participants. The second main level (medium length or longish aspect ratio), can be reached by either a parameter increase (upward jump) or decrease (downward jump), if only main levels are considered. First and third main level can be reached from only

Table 5.8.: Confusion matrix for the sign of a change in length or aspect ratio (+ increase, – decrease), for both materials separately, pooled over participants. Different sample sizes are due to the data series not being adjusted for equal numbers of increases and decreases.

		(a) length, glass		(b) length, wood	
		true		true	
		+	–	+	–
estimated	+	122	20	136	4
	–	7	107	3	135
		$Acc=0.89, \chi^2(1)=161.0$		$Acc=0.97, \chi^2(1)=250.7$	
		(c) aspect ratio, glass		(d) aspect ratio, wood	
		true		true	
		+	–	+	–
estimated	+	116	10	108	5
	–	21	125	15	122
		$Acc=0.89, \chi^2(1)=163.3$		$Acc=0.92, \chi^2(1)=177.4$	

one direction, but with different step size (one or two levels). For wood, the judgments of length following an increase were significantly larger than those following a decrease (Mann-Whitney U test, $p < 0.001$). This leads to a kind of hysteresis curve in this case. In all other cases, the difference between increase and decrease towards the unknown plate was not significant.

5.2.5.6. Number of discriminable levels

From the perspective of a sonification designer, it seems beneficial to know the number of absolute discriminable levels for each parameter. This obviously depends on the number of errors we tolerate for the given target application. Such a function can be derived on the basis of the effect size (see also Groß-Vogt et al. 2021 or Sec. 8.3.5). The estimated values for the main levels, shown in Fig. 5.8, do not follow a normal distribution, due to end effects and bimodal distribution of pooled data. As our main criteria for differentiating between the main levels is the mean value, we nevertheless use the traditional parametric definition of Cohen's d to measure effect sizes between main levels. For both steps between adjacent main levels, Cohen's d is calculated as the difference between mean values μ_A and μ_B divided by the pooled standard deviation s_p for independent

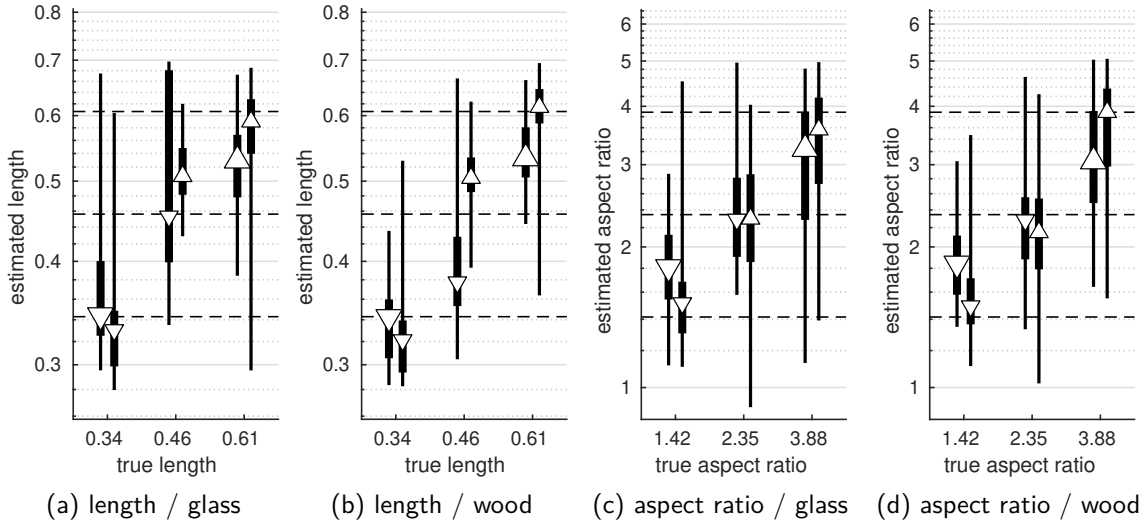


Figure 5.8.: Estimated value vs. true value for the 4 combinations of material and parameter. Triangles are median values, error bars represent 5- and 95- as well as 25- and 75-percentiles. \triangle = increase by 1 level, Δ = increase by 2 levels, ∇ = decrease by 1 level, and ∇ = decrease by 2 levels. Dashed lines are the true values.

samples of unequal size and variance:

$$d = \frac{\mu_B - \mu_A}{s_p} . \quad (5.27)$$

Values for Cohen's d are given in Tab. 5.9. Effect sizes of 0.1, 0.3, 0.5, 1, 1.2, and 2 are usually interpreted to signify a small, medium, strong, large, very large, and huge effect, respectively. Summing up both effect sizes, divided by the desired threshold effect size d_t , yields the number of steps that can be discriminated with the given d_t . Adding the one level that can always be heard yields the number of discriminable main levels, as was done by Groß-Vogt et al. (2021):

$$N = \frac{\sum_{i=1}^2 (d_{i,i+1})}{d_t} + 1 . \quad (5.28)$$

In addition, we can now insert any target effect size, and retrieve an estimate for the corresponding number of discriminable levels. Or vice versa.

Any value of effect size can also be expressed in terms of probability of superiority P_s , i.e., the probability that a larger true value leads to a larger estimated value. In case of only two values, it equals the probability that a participant's judgment is correct. P_s is actually the same as the area under the receiver operating characteristic curve (AUROC, or simply AUC), sometimes also called common

Table 5.9.: Cohen's d between neighboring main levels of length and aspect ratio, pooled over participants, for both materials, respectively.

		glass	wood
length	small vs. medium	1.83	1.63
	medium vs. large	0.07	1.44
aspect	compact vs. longish	0.95	0.68
	longish vs. bar-shaped	0.90	1.21

language effect size (CL):

$$P_s = \Phi \left(\frac{d}{\sqrt{2}} \right) , \quad (5.29)$$

with the cumulative standard normal distribution Φ and Cohen's d (Ruscio 2008). The other way around, Cohen's d can be expressed in form of P_s via

$$d = \sqrt{2}\Phi^{-1}(P_s) . \quad (5.30)$$

This lets us now compute the number of discriminable steps N for any desired probability of superiority, via Eq. 5.28. In addition, 95% confidence intervals of the effect sizes and thus number of discriminable levels can be computed. These are based on confidence intervals of the noncentrality parameters of the assumed t -distributions (Howell

5. Auditory perception and information capacity of rectangular plates

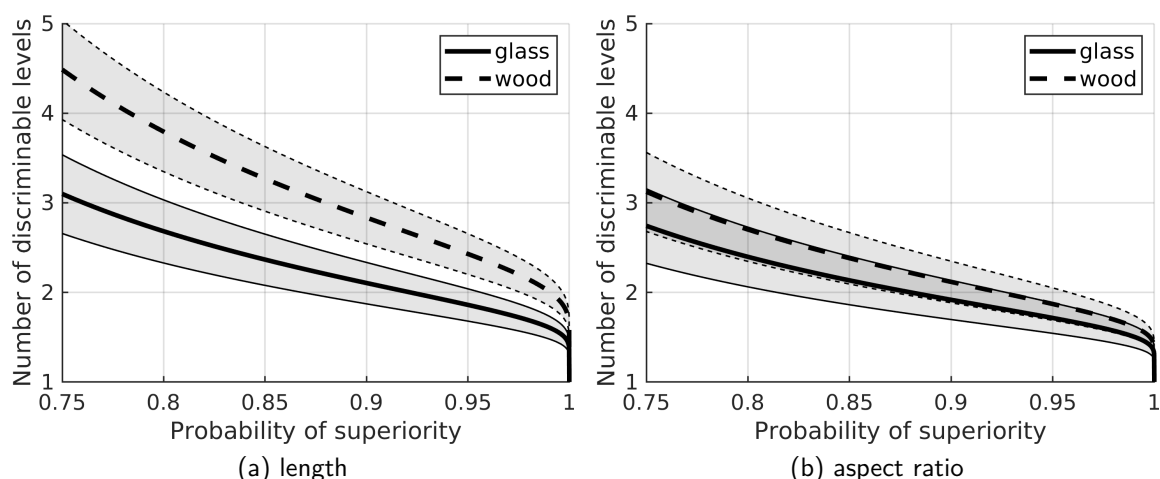


Figure 5.9.: The number of discriminable levels, plotted against the probability of superiority. The shaded areas are 95% confidence intervals.

Table 5.10.: Top 10 confused pairs of plates, together with their probability of confusion. Pairs include both combinations of reference and unknown plate.

$P(\text{conf.})$	confused plates	
0.24	medium / compact	large / longish
0.14	small / longish	medium / bar-shaped
0.14	medium / longish	large / bar-shaped
0.13	small / compact	medium / longish
0.12	small / compact	large / bar-shaped
0.12	small / compact	medium / bar-shaped
0.11	medium / compact	large / bar-shaped
0.08	small / compact	large / longish
0.07	small / longish	large / bar-shaped
0.07	medium / longish	large / compact

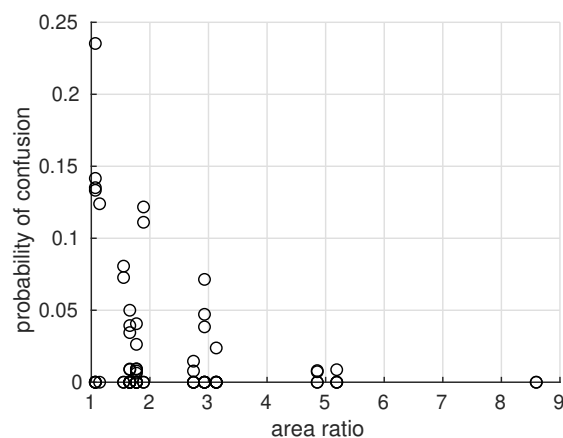


Figure 5.10.: Confusion probability vs. relative area difference for the 36 pairs of plates, based on jitter-corrected true and estimated values. Area differences are relative to the smaller plate's area, respectively; estimated values are rounded to main levels.

2011).⁶ The resulting estimate functions are shown in Fig. 5.9, separately for length and aspect ratio, and separately for glass and wood.

5.2.5.7. Confusion between individual shapes

The participants' answers can be interpreted as a confusion between the estimated plate and the true plate. If only main levels are considered, and estimated values are rounded to main values, each pair of true and answered plate is interpreted as a confused pair. Based on the frequency for each pair (the order of reference and unknown plate doesn't matter), a probability of confusion is constructed. The top 10 confused pairs of plates are listed in

Tab. 5.10.

They all exhibit a confusion probability larger than 5% and differ in area by less than factor 3 (see Fig. 5.10). The areas of all combinations of main levels of length and aspect ratio are given in Tab. 5.11. Figure 5.10 suggests that plates of equal area are likely to be confused. This may be attributed to the assumption that participants sometimes tend to answer in terms of area instead of the demanded parameter. Note that any pair of the four parameters length, width, aspect ratio, and area

⁶for implementation, see MBESS package for R: <https://rdrr.io/cran/MBESS/man/conf.limits.nct.html>

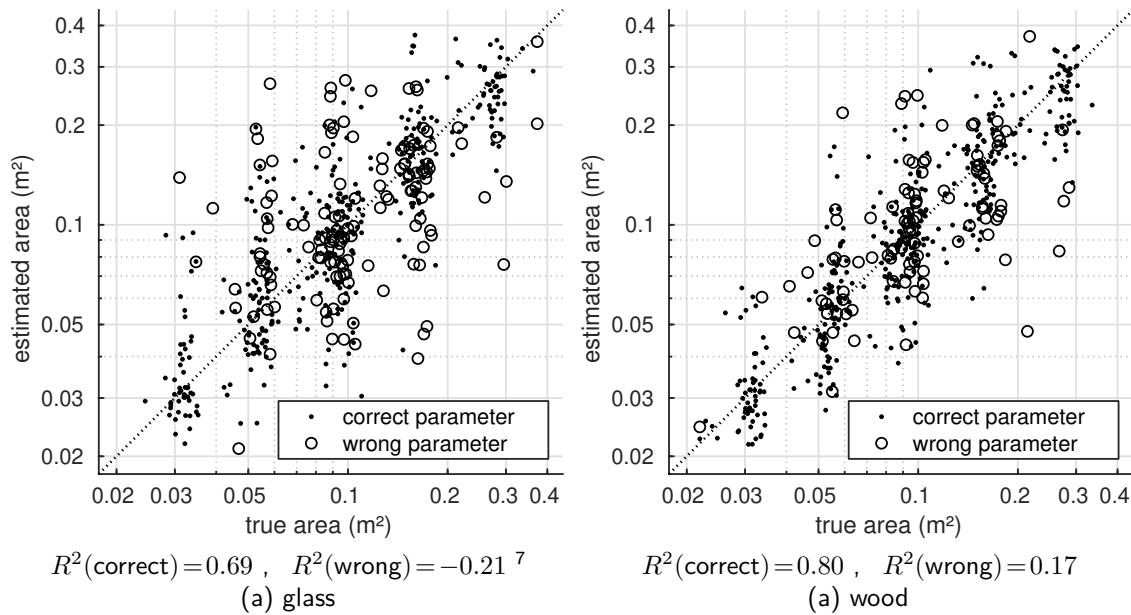


Figure 5.11.: Estimated area vs. true area of all stimuli and all participants, (a) for the glass plate and (b) the wooden plate. Coefficients of determination R^2 relate to the ideal (dotted line), for estimations in the correct and wrong parameter dimension separately (on a logarithmic scale).

Table 5.11.: Surface areas of the rendered plates in m^2 for all combinations of the main levels of length and aspect ratio.

	small	medium	large
compact	0.083	0.147	0.260
longish	0.050	0.089	0.157
bar-shaped	0.030	0.054	0.095

is sufficient to describe the 2D plate dimensions.

Every combination of length and aspect ratio describes a surface area. Even if participants were not explicitly asked about area, for each true area of the modeled plate, an estimated area can thus be calculated based on a participant’s estimated length and aspect ratio. Figure 5.11 shows the estimated areas of all plates in relation to their true area. Note that these include all individual answers by all participants. The coefficients of determination confirm the visual impression that judgments in the correct parameter dimension lead to much stronger agreement with the true area than those judgments where participants changed the wrong parameter.

⁷A negative value of R^2 in this case means that the true area performs worse in predicting the estimated area than the geometric mean of true areas would.

5.2.5.8. Individual participants’ interaction strategies

Although all participants received the same introductions, including the recommendation to tap everywhere within the tapping region, they developed quite different tapping strategies to explore the rendered plates. Figure 5.12 shows the impact patterns of all 14 participants. Impacts were detected via the tracking system as local minima below a threshold of 5 mm distance to the plate. This high threshold is necessary due to the low spatial (1 mm) and temporal (50 ms) resolution of the recorded data. For participant P7, the recording of tracking data failed; the impact pattern was therefore derived from a scan of the actual bumps in the overlay paper.

Especially interesting are the patterns of P6, P8, and P11, who independently from each other concentrated on three distinct spots for tapping. This behavior was possibly induced by the introductory task where participants had to tap on three marked spots on a physical wooden and glass plate, respectively. According to their patterns, they interacted mainly within the principal axis in x - and y -direction and thus excited only half of the modes. As their performance was pretty average (see Fig. 5.7), no conclusions can be drawn about possible benefits or drawbacks of this strategy.

5. Auditory perception and information capacity of rectangular plates

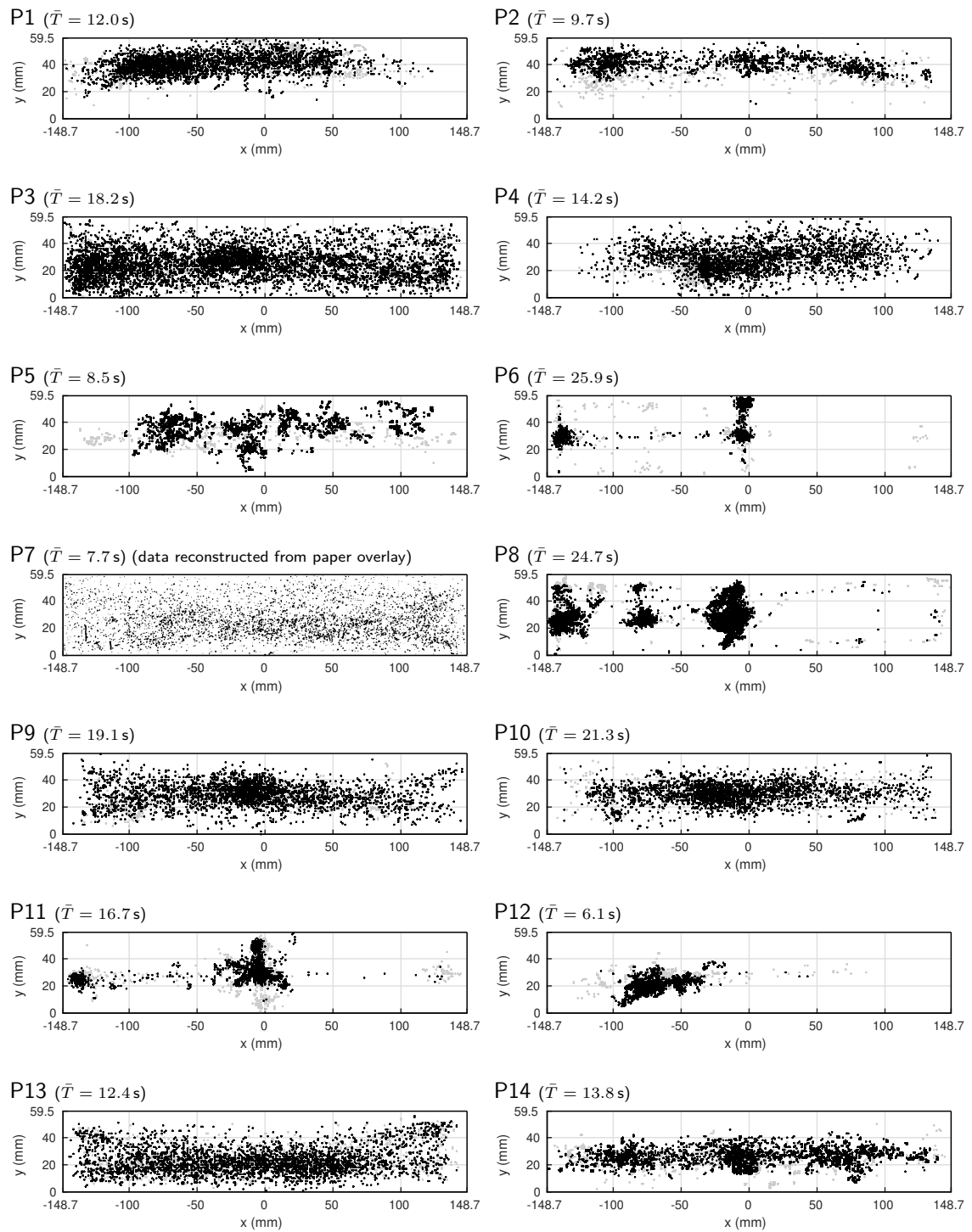


Figure 5.12.: Impact patterns of all individual participants (gray = training, black = test), together with the average response time per trial \bar{T} .

On average, participants took 15.0s to form an answer, measured from the time they first switched to the unknown plate. There was no significant difference between glass and wood (15.7s and 14.3s, respectively). No significant correlation was found between response time and accuracy.

5.2.6. Discussion: why is it so difficult?

The two parameters length and aspect ratio are mainly connected to absolute frequency factor (and thus pitch) and relative frequency ratios between modes (and thus intervals and modal density), respectively.

The range of length (between large and small) corresponds to a frequency ratio of 0.607/0.343 or 1.77 for all modes (at constant aspect ratio). The range of aspect ratio defines a ratio of frequency ratios, or, in other words, the frequency ratio between modes $n/0$ and $0/n$. The ratio between bar-shaped and compact was 3.884/1.416 or 2.74.

According to informal listening tests, the pitch of the stimuli used in the experiment can be approximated as the frequency of mode 3/0 which is the 2nd-lowest mode that is constant over aspect ratio. In case of glass, it ranges from 438 Hz (large) to 1372 Hz (small); for wood, this range reaches from 424 Hz to 1329 Hz.

The JND in base frequency has been derived by Lutfi and Stoelinga (2010) (see Sec. 2.3.4). If mode 3/0 is taken as base frequency, then the average frequency range between 431 Hz and 1351 Hz (averaged across materials) between large and small length takes about 558 JNDs stacked on top of each other.

As some of the lower resonant frequencies are more than 10% apart from each other, they can be perceived as individual pitches (Thurlow and Bernstein 1957). According to the JND derived by Fantini and Viemeister (1987) (see Sec. 2.3.5), if JNDs are stacked one upon another, the range of aspect ratio holds about $\log_{1.012}(2.74) = 84.5$ JNDs.

The statistical measure of modal density provides another cue for the aspect ratio. We assume the JNDs derived by Stoelinga and Lutfi (2011) (see Sec. 2.3.6). In case of glass, the modal densities of the four extreme combinations are 0.00148 (small/bar-shaped), 0.00344 (small/compact), 0.00407 (large/bar-shaped), and 0.00927 (large/compact). For wood, these are 0.00228 (small/bar-shaped), 0.00536 (small/com-

compact), 0.00599 (large/bar-shaped), and 0.01166 (large/compact). The ratio between modal densities of bar-shaped and compact aspect ratios is almost constant across lengths and materials, on average 2.22. As the JND is also assumed to be a relative measure, the range ratio of modal densities that is given through the ratio of aspect ratios fits about $\log_{1.3}(2.22) = 3.0$ JNDs.

Figure 5.13 visualizes the frequency ratios (relative to the frequency of mode 2/0 for a quadratic plate) as a function of aspect ratio, for isotropic (glass) and orthotropic materials (wood) with $\Omega=2$. If the lowest two modes cross each other, their frequency ratio is no longer a reliable descriptor. In case of glass ($\Omega=1$), modes 1/1 and 2/0 cross just above the lowest main level (compact). In case of wood ($\Omega=2$), this effect is even more dramatic, as orthotropy basically works against aspect ratio, so that frequency ratios are ambiguous for a majority of the parameter range. The effective aspect ratio is the original aspect ratio multiplied by the orthotropy factor. Based on these considerations, it is quite surprising, that accuracies for discriminating main levels as well as directions of aspect ratio did not significantly differ between wood and glass.

A by-product of the crossing modes is that perceived pitch may differ significantly depending on the aspect ratio, especially for the wooden plate, leading to a confound with length.

One may argue that the lowest modes are anyway barely radiated, especially for the bar-shaped plates, so that only frequency ratios of higher modes can be evaluated for estimating the aspect ratio. The radiation efficiency can be roughly approximated by a 1st-order high-pass filter with cutoff at the critical frequency. f_{cr} is independent of length and aspect ratio, and equals 1162 Hz for glass and 1763 Hz for wood (with $c_0 = 344$ m/s and $D = \sqrt{D_1 D_3}$). For comparison, in case of glass, the frequency of mode 3/0 (the 2nd-lowest mode of constant frequency over aspect ratios) varies between 438 Hz for large and 1372 Hz for small length. In case of wood, it is even lower with 234 Hz for large and 733 Hz for small length. In consequence, participants are barely able to actually utilize interval relationships of single modes for estimating the aspect ratio of an unknown plate. The slightly better performance in parameter identification that was achieved for the wooden plates might even be attributed to this filtering of lower modes which confound the perceptual descriptors for length and aspect ratio.

The difference in surface hardness is quite prominent between glass and wood. While glass is mod-

5. Auditory perception and information capacity of rectangular plates

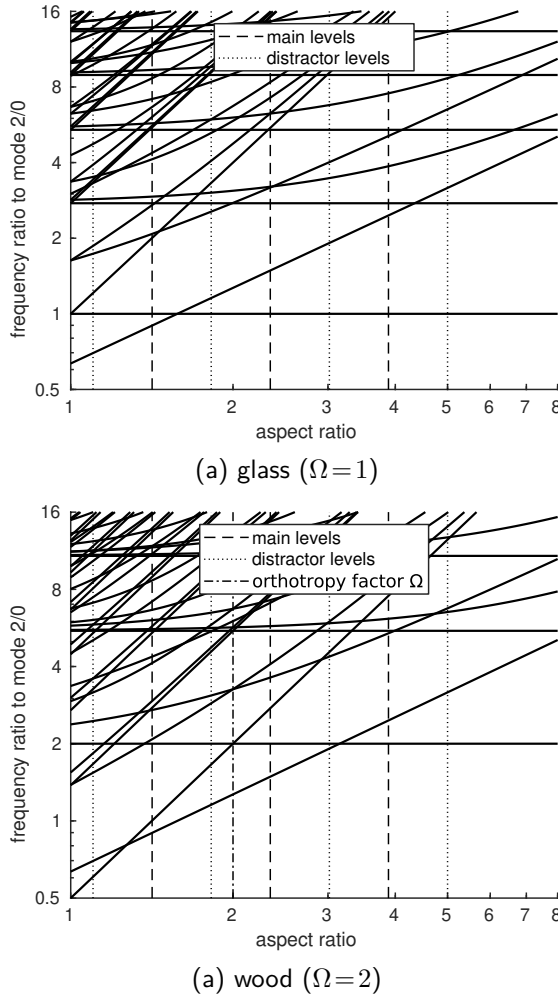


Figure 5.13.: Frequency ratios, relative to the frequency of mode 2/0 for a quadratic plate, as a function of aspect ratio, for both materials or orthotropy factors.

eled with a Brinell hardness of 1550 kgf, leading to an upper cutoff frequency 45 kHz—irrelevant for sound perception, the low hardness of wood (1.3 kgf) leads to an upper cutoff frequency of 2410 Hz for the model plate (see Sec. 3.2.9). Above this upper cutoff frequency, the sound pressure level drops drastically.

We can conclude that the participants had difficulties in performing the demanded task due to several reasons. First, the sound parameters that conveyed the physical information of length and aspect ratio were not entirely independent. We assume that participants based their judgments of length on either base frequency or pitch. Some of the lowest modes, however, are not radiated,

depending on the aspect ratio, which leads to interference between length and aspect ratio. For both materials, the critical frequency is so high (wood: $f_{cr} = 1752$ Hz, glass: $f_{cr} = 1155$ Hz) that lower modes could have played only a minor role in the perception of length. While inaudible lower partials can be reconstructed computationally by matching a theoretical model of a rectangular plate, as has been shown in Sec. 5.1, participants failed to evaluate this information. For aspect ratio, it seems likely that the primary source of information that participants exploited was modal density. At $P_s = 0.75$ (a common value for the JND), about 3 aspect ratios could be discriminated, which is equals the number of JNDs of modal density which fit the given parameter range. Follow-up experiments are therefore necessary to examine how auditory augmentations of rectangular plates need to be tuned in order to maximize the conveyed physical information.

5.3. Auditory discrimination of material and aspect ratio

The web-based implementation of this experiment was done by Aurenhammer (2021) as part of his bachelor thesis under my supervision, based on my original ideas.

We have learned from the first experiment that size and aspect ratio can indeed be employed as carrier parameters of a 2D auditory display. We also learned, however, that aspect ratio and orthotropy confound each other and that their parameter range should be carefully chosen to avoid crossing partials as far as possible. According to Fig. 5.13a, a low limit of aspect ratio at $r_a = 2$ seems appropriate. The range, however, might be extended up to a more bar-like shape. Instead of the length (i.e., surface area) from experiment 1, we want to try another approach. From the literature we know that humans are quite good at identifying gross density categories of materials such as glass/metal in contrast to plastic/wood (see Sec. 2.2.1) and that even within-category discrimination (metal vs. non-metal) is somehow possible (see Sec. 2.2.3). It may therefore be possible to span a salient 2D parameter space by density and metallicity. As density mainly affects pitch, it seems more convenient to rather employ a more high-level meta-parameter such as longitudinal wave velocity c_L for this purpose—we will name it rigidity to underline its physical meaning. Due to the common effect on pitch we therefore use a constant length in this experiment. The parameter

dimensions of the auditory display are thus rigidity, metallicity, and aspect ratio.

While in experiment 1 the participants actively explored unknown model plates through an physical interface (Sec. 5.2), experiment 2 tested pure auditory perception of pre-rendered sounds based on the same physical model implementation.

5.3.1. Stimuli

The model plates are described by 3 meta-parameters: rigidity, metallicity, and aspect ratio. These affect the resulting sound in multiple ways; however, they can be roughly attributed to the sound parameters base frequency, frequency-dependent damping law, and modal density or intervals between partials, respectively. As all these are based on either frequency or duration, the model parameters are scaled accordingly to provide an exponential scale that roughly corresponds to their perception (see Sec. 2.3). While all those parameters affect the plate on a continuous interval scale, only discrete levels are used during the experiment. Metallicity takes 2 levels (non-metallic, metallic). Aspect ratio takes 3 levels (compact, longish, and bar-shaped). Rigidity takes 3 levels, with labels for individual material categories that depend on the state of metallicity. For non-metals, these are plastic, wood, and glass. For metals, these are gold, brass, and aluminium. The metallicity parameter blends between the corresponding pair of non-metal and metal material which otherwise share almost similar pitch.

For blending between non-metal and metal materials (e.g., glass and aluminium), we need to find several pairs of materials with equal longitudinal wave velocity c_L (and thus base frequency). The algorithm for robotic perception (Sec. 5.1) revealed that c_L can be only perceived in combination with thickness h . Their product hc_L in frequency factor $\hat{\Phi}$ (see Eq. 3.54 in Sec. 3.2.5) is not separable from the resulting sound, at least not by our modal synthesis model from Ch. 3. We therefore take advantage of the freely adjustable thickness to align selected material categories with their corresponding value of c_L to an equally-spaced grid of perfectly tuned and matched base frequencies. The glass plate is selected as reference with a plausible thickness of 8 mm for the intended use as a table. Its frequency factor and thus pitch is thus already defined by the physical parameters. The two rigidity levels below (and also those of the metals) are tuned to frequency factors of 0.75 and 1.5 octaves below by small adjustments of h . Figure 5.14 plots

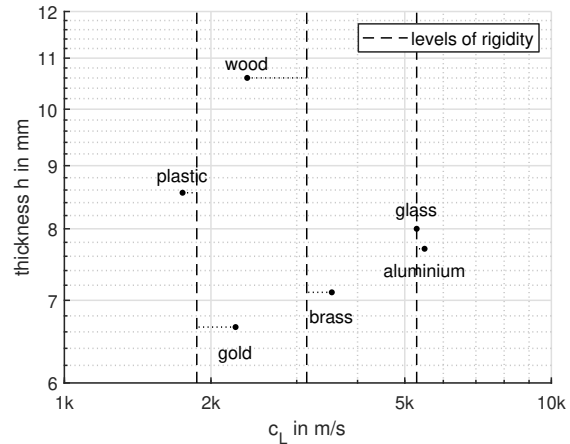


Figure 5.14.: The adjusted thickness h plotted against the longitudinal wave velocity c_L for all 6 materials used in the experiment. Dashed lines mark the 3 levels of rigidity that are effectively achieved by adjusting the plate's thickness.

the adjusted thickness against the longitudinal wave velocity c_L for all 6 materials used in the experiment.

The length of the model plate is set to a constant value of 0.42 m which is a compromise between a plausible size and the desired base frequency being not too low. The aspect ratio (length:width) ranges from 2:1 (compact) via 4:1 (longish) to 8:1 (bar-shaped). The virtual table thus transforms from a small but still usable size down to a bar of same length. A plausible physical interpretation might be that the table is either made of one single plate (2:1) (normal table) or many narrow planks (8:1) (garden table), or anything in between. The effect on the frequency factors of the modes is shown in Fig. 5.15. The lowest two modes 1/1 and 2/0 are usually barely radiated. The chosen levels for experiment 2 are marked; the parameter range is a bit larger than in experiment 1 (see also Fig. 5.13).

According to informal listening sessions, the 2nd lowest constant mode 3/0 has a strong effect on the perceived pitch. The excitation position is therefore set constant, on the edge of the plate, so that as many modes are excited as possible⁸, but in the maximum of this mode, at normalized position [0.3083, 0]. It is a good compromise between attenuation of 1/1 and 2/0 on the one hand (which would pull pitch down if radiated in case of a compact plate), and boosting 3/0 on the other hand, see Fig. 5.16.

Metallicity cross-fades between viscoelastic and

⁸Note that free boundary conditions are used on all edges.

5. Auditory perception and information capacity of rectangular plates

Table 5.12.: Model coefficients of the rendered plates. The thermoelastic constants of metal material are equally used for non-metals.

		non-metal			metal			
		plastic	wood	glass	gold	brass	aluminum	
thickness	h	8.557	10.602	8.000	6.659	7.105	7.707	mm
density	ρ	1150	590	2550	19300	8500	2700	kg m^{-3}
Young's modulus	E	3.20	3.29	66.90	80.00	95.00	72.00	GPa
Poisson's ratio	ν	0.300	0.100	0.250	0.423	0.330	0.340	
hardness	HB	34	15	1550	22	100	36	$\text{kgf} = 9.80665 \text{ N}$
thermoelastic constants	R_{1t}	—	—	—	64.31	22.42	24.84	$10^{-3} \text{ rad m}^2 \text{ s}^{-1}$
	c_{1t}	—	—	—	1.251	0.489	0.977	$10^{-3} \text{ rad s}^{-1}$
metallicity	H	0						1
viscoelastic loss factor	η_N	5.7/ c_L			0.57/ c_L			
wave velocity	c_L	1748.7	2373.6	5290.0	2246.9	3541.5	5491.1	m s^{-1}
length	l_x	0.420						m
aspect ratio	r_a	{2, 4, 8}						

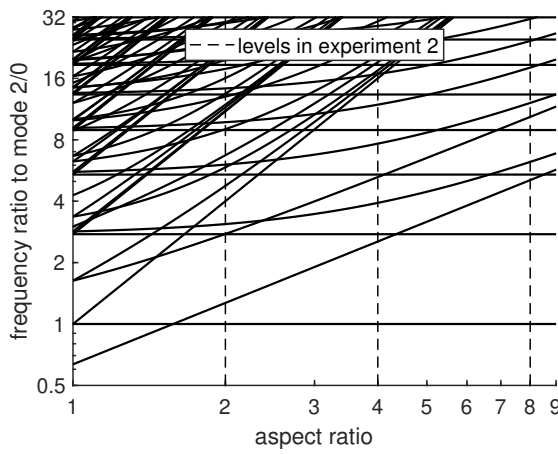


Figure 5.15.: Frequency factors of the partials of a rectangular plate (relative to the lowest frequency at aspect ratio 1:1) as a function of aspect ratio. Dashed vertical lines depict the levels used in experiment 2.

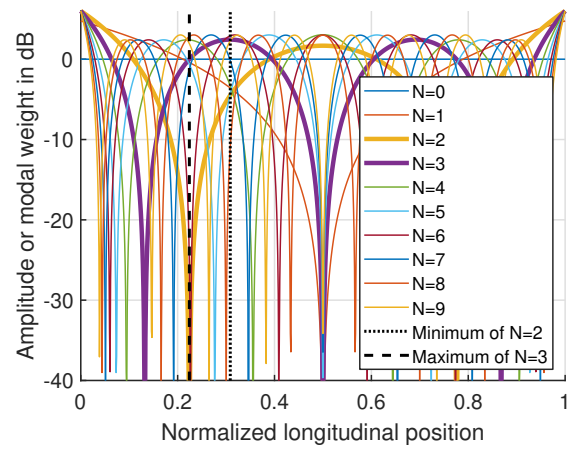


Figure 5.16.: Modal weights of a vibrating bar with free ends as a function of excitation position.

thermoelastic damping, as its value between 0 and 1 is directly taken as H . While viscoelastic decay factors α_v are proportional to frequency, the decay factors due to thermoelastic damping are constant over all frequencies but weighted for each mode individually, depending on the mode shapes, as shown in Sec. 3.2.6. Within the experiment, metallicity takes only the extreme values 0 (non-metallic) and 1 (metallic). In addition to H , metallicity blends between the non-metallic and metallic material constants of the 3 pairs of materials on an exponential scale, which means that the linear input range between 0 and 1 is mapped to an exponential output range. The employed material constants are given

in Tab. 5.12. The thermoelastic constants R_{1t} and c_{1t} are pre-computed via the underlying physical constants based on Eq. 3.81 and 3.79.

Additional damping is introduced to create a plausible overall decay time, different to the ideal free plate that is assumed to be freely hovering in the air. This is achieved by a constant bias to the decay factors. We argue that almost any frequency-dependent damping that is higher than that of the ideal free plate can be achieved by some kind of physical physical suspension. In other words, it seems physically feasible to set any desired decay time as long as it is lower than that of the free plate. The decay factors of all rendered materials are therefore shifted so that the 3/0 mode matches a desired decay time that can be chosen freely. Mode 3/0 roughly controls pitch and is usually the longest decaying mode if those below are neglected due to

excitation in its maximum. Within the experiment, two decay times T_{60} of 0.15 s and 0.45 s are used in two different conditions.

Stimuli are pre-rendered by using the complete physical model as described in Ch. 3. As excitation signal for a single impact, a Hann window of constant length 0.5 ms is used. The duration is chosen to perceptually match the pen excitation through a paper overlay that was the case in experiment 1. Each stimulus consists of the same model plate that is excited by 4 successive impacts in time intervals of 200 ms. A less machine-like, more natural rhythm is achieved by randomization of the exact time onset by $\pm 5\%$ of the interval or ± 10 ms.

All combinations of levels of metallicity, rigidity, and aspect ratio lead to 18 model plates for each of the two damping conditions, or 36 different stimuli in total. To prevent participants from directly memorizing individual sounds, the 3 meta-parameters (all in the range between 0 and 1) are jittered uniformly by ± 0.05 , as done in comparable studies (e.g., Lutfi and Liu 2007). In addition, each individual impact of the excitation is jittered by ± 3 dB in amplitude and $\pm 10\%$ in duration.

Each of the 36 stimuli is rendered in 4 different variations, two for training and two for the two repetitions in the test.

5.3.2. Apparatus and procedure

The experiment was implemented in form of a web page, with the help of the open-source JavaScript library jsPsych⁹ (Leeuw 2015) which is designed for running cognitive experiments online in the web browser. The experiment was hosted free of charge on the related Cognition¹⁰ platform. The source code of the experiment, the Matlab script for generating the stimuli, as well as the rendered sounds are available in Source 5.1.

After a brief landing page which summarized conditions, procedure, and the motivation behind the experiment, participants navigated through the experiment by using their mouse or other pointing device. First, they had to indicate if they would classify themselves as trained listeners (e.g., due to musical training or professional background). Participants were then requested to put on the best headphones available. During a passive introductory phase, all parameter dimensions were explained individually with the help of sound examples. During an

Figure 5.17.: The testing page for a single stimulus.

active training phase, participants had the possibility to compare all 18 parameter combinations within one damping condition by clicking on the appropriate buttons of two 9×9 matrices for non-metals and metals, respectively. They were free to decide when ready to start the test for this condition. The two damping conditions (each consisting of active training and subsequent test) were presented one after another, in randomized order across participants. Participants were encouraged to perform the experiment without longer pauses; the possibility to take a break was announced between the two damping conditions. During the testing phase, they had to identify the given stimuli in all three parameter dimensions, with the possibility of infinite replays, as shown in Fig. 5.17. One damping condition contained two repetitions of each of the 18 stimuli in random order. A progress bar indicated the overall progress of the experiment.

5.3.3. Participants

The experiment was announced on newsgroups and social media platforms that are associated to the local students of sound engineering. In addition, colleagues and personal acquaintances were invited individually. In order to encourage people to participate, a lottery was initiated among participants who completed the experiment, with the opportunity to win one of two vouchers for local shops (50 EUR each, for books or professional audio equipment

⁹jsPsych JavaScript library: <https://www.jspsych.org/>

¹⁰Cognition platform: <https://www.cognition.run>

5. Auditory perception and information capacity of rectangular plates

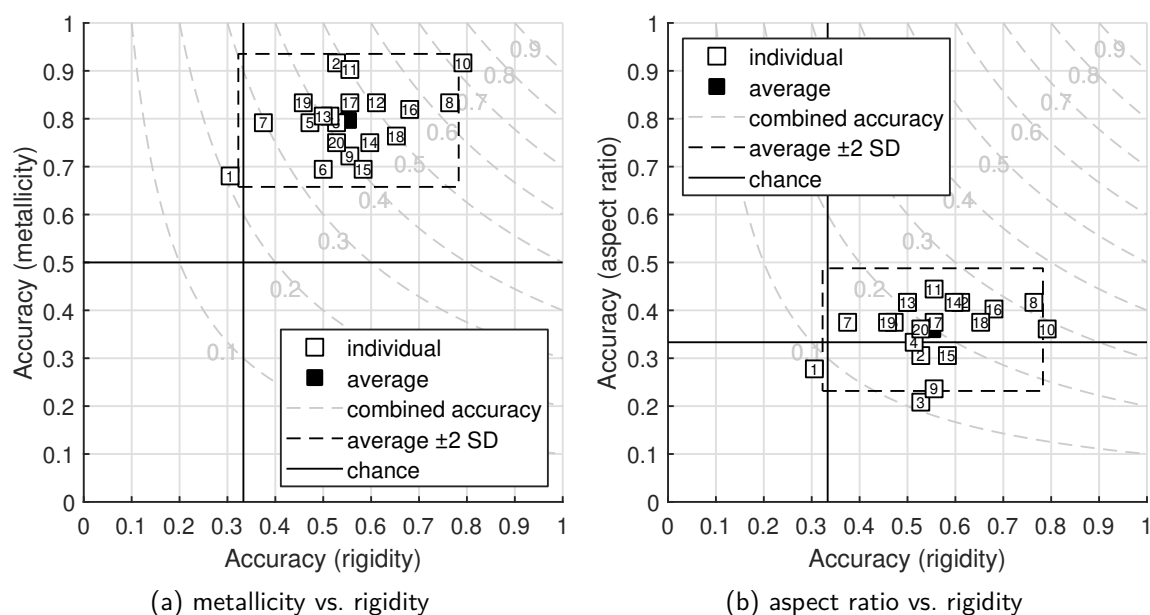


Figure 5.18.: Individual participants' accuracies together with their average, for the three parameters (pooled over dampings, repetitions, and both other parameters).

alternatively). In total, 20 anonymous participants finished the experiment. 14 of these classified themselves as trained listeners.

5.3.4. Results

The results from the experiment were collected in CSV format by jsPsych. Based on these files, the statistical analysis was done in Matlab.

5.3.4.1. Overview

In total, 16% of the stimuli were correctly identified in all three parameter dimensions (pooled over participants and dampings, with 95% confidence interval CI_{95} between 13% and 18%). There was no significant effect of damping on this result. While this seems incredibly low, the chance level of $1/18 = 5.6\%$ in this case must be considered. As the number of trials was balanced across classes, this simple metric of percent correct classifications, i.e., the accuracy, is an appropriate measure of the participants' performance.

For comparison between the three parameter dimensions, their individual accuracy was computed. For identification of metallicity (i.e., non-metal vs. metal), the overall accuracy was 0.80 ($CI_{95} = [0.78, 0.82]$), pooled over participants, dampings, rigidities, and aspect ratios. For rigid-

ity identification (i.e., plastic/gold vs. wood/brass vs. glass/aluminum), the overall accuracy was 0.55 ($CI_{95} = [0.53, 0.58]$), pooled over participants, dampings, metallicities, and aspect ratios. For aspect ratio identification (i.e., compact vs. longish vs. bar-shaped), the overall accuracy was 0.36 ($CI_{95} = [0.33, 0.38]$), pooled over participants, dampings, metallicities, and aspect ratios.

One-tailed Wilcoxon signed-ranks tests were used for pairwise comparisons between adjacent levels within the parameters rigidity and aspect ratio. Response data could take on ordinal values between 1 and 3, equaling the distinct levels. A 5% threshold for statistical significance was used. For rigidity, level 2 (wood/brass) was rated significantly higher than level 1 (plastic/gold) ($Z = 30142, p < 0.001$), and level 3 (glass/aluminum) was rated significantly higher than level 1 (wood/brass) ($Z = 50806, p < 0.001$). For aspect ratio, level 2 (longish) was rated significantly higher than level 1 (compact) ($Z = 31539, p < 0.001$), but level 3 (bar-shaped) was not rated significantly higher than level 2 (longish) ($Z = 28996, p = 0.244$). Metallicity included only two categorical levels. Fisher's exact test revealed that metals were rated significantly different to non-metals ($p < 0.001$, odds ratio: 15.6).

For each participant, an individual accuracy was computed for each parameter dimension. In each case, the data was pooled over the two other param-

Table 5.13.: Top 16 confused pairs of plates, together with their probability of confusion. Pairs include both combinations of true and estimated plate.

$P(\text{conf.})$	confused plates	
0.331	wood / bar	wood / longish
0.306	glass / bar	glass / longish
0.212	aluminum / bar	glass / longish
0.206	plastic / bar	plastic / longish
0.175	brass / bar	brass / longish
0.169	plastic / longish	plastic / compact
0.163	brass / bar	gold / bar
0.156	brass / compact	gold / compact
0.156	aluminum / bar	aluminum / longish
0.150	plastic / bar	plastic / compact
0.150	plastic / bar	wood / longish
0.150	aluminum / bar	glass / bar
0.144	aluminum / compact	brass / compact
0.144	gold / bar	gold / compact
0.138	gold / longish	brass / compact
0.131	gold / bar	gold / longish

eters as well as over the two damping conditions and repetitions. These accuracies are plotted against each other in Fig. 5.18, together with the average over all participants. Standard deviations across participants were 0.07 for metallicity, 0.12 for rigidity, and 0.06 for aspect ratio.

From visual inspection, no participant seems to represent an obvious outlier. While a few participants are more than two standard deviations away from the average for material (participants 1 and 10) and aspect ratio (participant 3), they generally spread rather symmetrically around the average.

Neither the average trial duration nor the average number of replays had a significant effect on the participants' average accuracy. The overall accuracy, averaged over participants, was equal for trained and untrained participants (0.61 vs. 0.57). The difference was strongest for rigidity discrimination (0.58 vs. 0.50), but not significant ($t(18) = 1.65$, $p = 0.058$).

In some special cases, some plates may sound similar, even if their physical parameters are very different. For example, a large but dense plate may exhibit the same modal frequencies as a small plate of low density. Even though the physical parameters in this experiment have already been tuned based on previous results, we are interested in those pairs of plates which were difficult to discriminate. For each combination of two stimuli (the order doesn't matter), the probability that one was identified as the other was computed. This confusion probability is shown in Tab. 5.13 for the 16 most confused

Table 5.14.: Confusion matrix for metallicity, for both damping conditions separately, pooled over participants, rigidities and aspect ratios.

(a) weak damping

		true	
		non-metal	metal
estimated	non-metal	299	67
	metal	61	293

$Acc = 0.82$, $\chi^2(1) = 299.1$

(b) strong damping

		true	
		non-metal	metal
estimated	non-metal	293	98
	metal	67	262

$Acc = 0.77$, $\chi^2(1) = 212.8$

pairs. All of them had confusion probabilities above 0.13, while all less frequent confusions had probabilities below 0.12. In summary, 12 of the top 16 can be attributed to confusions between aspect ratios (top 5 occupied by confusion between longish and bar-shaped), 12 involve a bar-shaped plate, 5 refer to rigidity confusions (3× brass/gold, 1× aluminum/brass, 1× wood/plastic), and 2 refer to metallicity confusions between aluminum and glass.

5.3.4.2. Confusion between non-metals and metals

The overall confusion between metals and non-metals is shown in Tab. 5.14 for both dampings separately. The data was pooled over participants, rigidities, and aspect ratios. Participants achieved a slightly (but significantly) higher accuracy for weakly damped plates ($Acc = 0.82$, $CI_{95} = [0.79, 0.85]$) than for strongly damped plates ($Acc = 0.77$, $CI_{95} = [0.74, 0.80]$). This is mainly attributed to false negatives for metal in case of strong damping.

In Tab. 5.15, the data is pooled over both dampings, but split into two groups of rigidity, in order to reveal influences of rigidity on perceived metallicity. While confusion was equally high for plastic/gold and wood/brass ($Acc = 0.91$), it was significantly lower for glass/aluminum ($Acc = 0.57$, $CI_{95} = [0.53, 0.62]$). Even in the latter case, however, the results were still significantly different from chance ($\chi^2(1) = 9.9$, $p = 0.002$).

A similar effect could be observed for the different aspect ratios, as shown in Tab. 5.16, pooled

5. Auditory perception and information capacity of rectangular plates

Table 5.15.: Confusion matrix for metallicity, for rigidities separately, pooled over participants, dampings, and aspect ratios.

(a) plastic/gold & wood/brass

		true	
		non-metal	metal
estimated	non-metal	436	43
	metal	44	437

$Acc=0.91, \chi^2(1)=643.5$

(b) glass/aluminum

		true	
		non-metal	metal
estimated	non-metal	156	122
	metal	84	118

$Acc=0.57, \chi^2(1)=9.9$

Table 5.16.: Confusion matrix for metallicity, for aspect ratios separately, pooled over participants, dampings, and rigidities.

(a) compact & longish

		true	
		non-metal	metal
estimated	non-metal	379	76
	metal	101	404

$Acc=0.82, \chi^2(1)=383.6$

(b) bar-shaped

		true	
		non-metal	metal
estimated	non-metal	213	89
	metal	27	151

$Acc=0.76, \chi^2(1)=137.3$

over both dampings, but split into two groups of aspect ratios, in order to reveal influences of aspect ratio on perceived metallicity. While accuracy was equally high for compact and longish plates (0.81 and 0.82, respectively), it was significantly lower for bar-shaped plates ($Acc=0.76, CI_{95}=[0.72, 0.80]$).

5.3.4.3. Confusion between rigidities

The total confusion between rigidities is summarized in Tab. 5.17, pooled over participants, metallicities, and aspect ratios, but separately for weak and strong damping. However, there was no significant differ-

Table 5.17.: Confusion between rigidities, for both dampings separately, pooled over participants, metallicities, and aspect ratios.

(a) weak damping

		true		
		low	medium	high
estimated	low	126	72	35
	medium	84	114	39
	high	30	54	166

$Acc=0.56, \chi^2(4)=216.5$

(b) strong damping

		true		
		low	medium	high
estimated	low	123	74	35
	medium	75	115	53
	high	42	51	152

$Acc=0.54, \chi^2(4)=166.0$

Table 5.18.: Confusion between rigidities for non-metals and metals separately, pooled over participants, dampings, and aspect ratios.

(a) non-metals

		true		
		plastic	wood	glass
estimated	plastic	142	65	44
	wood	69	143	35
	glass	29	32	161

$Acc=0.62, \chi^2(4)=291.2$

(b) metals

		true		
		gold	brass	aluminum
estimated	gold	107	81	26
	brass	90	86	57
	aluminum	43	73	157

$Acc=0.49, \chi^2(4)=133.1$

ence between dampings, with average accuracy of 0.55 ($CI_{95}=[0.53, 0.58]$).

As Tab. 5.18 shows, participants were significantly better in discriminating non-metals ($Acc=0.62, CI_{95}=[0.58, 0.65]$) than metals ($Acc=0.49, CI_{95}=[0.45, 0.52]$).

There was also a noticeable effect of aspect ratio on the accuracy in discriminating between rigidities. Table 5.19 shows confusion between rigidities for

Table 5.19.: Confusion between rigidities for the three aspect ratios, pooled over participants, dampings, and metallicities.

(a) compact

		true		
		low	medium	high
estimated	low	100	61	38
	medium	41	75	52
	high	19	24	70

$Acc=0.51, \chi^2(4)=82.3$

(b) longish & bar-shaped

		true		
		low	medium	high
estimated	low	149	85	32
	medium	118	154	40
	high	53	81	248

$Acc=0.57, \chi^2(4)=317.3$

the three aspect ratio separately, pooled over participants, dampings, and metallicities. While discrimination between rigidities was equally high for longish ($Acc=0.58$) and bar-shaped ($Acc=0.56$) plates, accuracy was lower for compact plates ($Acc=0.51$, $CI_{95}=[0.47, 0.56]$).

5.3.4.4. Confusion between aspect ratios

Overall discrimination between aspect ratios is depicted in Tab. 5.20 for both dampings, pooled over participants, rigidities, and metallicities. There was no significant difference in accuracies between weak and strong damping; average accuracy was 0.36 ($CI_{95}=[0.33, 0.38]$). Despite the low accuracy, both confusion matrices were significantly different from chance (weak damping: $\chi^2(4)=82.7$; strong damping: $\chi^2(4)=32.3$; both $p \leq 0.001$).

A significant effect of rigidity on the discrimination between aspect ratios was observed. While accuracy was similar for compact and longish plates (on average: 0.38, $CI_{95}=[0.35, 0.41]$), it was significantly lower—and not even significantly different from chance—for bar-shaped plates (0.32, $CI_{95}=[0.28, 0.36]$). This is depicted in Tab. 5.21, with data pooled over participants, dampings, and metallicities.

No significant effect of metallicity was observed, with overall accuracy of 0.36 ($CI_{95}=[0.33, 0.40]$) for non-metals and 0.36 ($CI_{95}=[0.32, 0.39]$) for metals.

Table 5.20.: Confusion between aspect ratios, for both dampings separately, pooled over participants, metallicities, and rigidities.

(a) weak damping

		true		
		compact	longish	bar
estimated	compact	116	72	46
	longish	39	59	116
	bar	85	109	78

$Acc=0.35, \chi^2(4)=82.7$

(b) strong damping

		true		
		compact	longish	bar
estimated	compact	113	81	59
	longish	51	58	87
	bar	76	101	94

$Acc=0.37, \chi^2(4)=32.3$

Table 5.21.: Confusion between aspect ratios, dependent of rigidity, pooled over participants, metallicities, and dampings.

(a) plastic/gold & wood/brass

		true		
		compact	longish	bar
estimated	compact	168	107	75
	longish	59	70	118
	bar	93	144	127

$Acc=0.36, \chi^2(4)=73.6$

(b) glass/aluminum

		true		
		compact	longish	bar
estimated	compact	60	47	30
	longish	31	47	85
	bar	68	66	45

$Acc=0.32, \chi^2(4)=44.2$

5.3.4.5. Number of discriminable levels

Other than in experiment 1, the response data is only ordinal. Nevertheless, we can get a rough estimate of the number of discriminable levels as a function of the probability of superiority, based on Cohen's d , as was done in experiment 1. The result is shown in Fig. 5.19.

In addition, we have one reference data point for

5. Auditory perception and information capacity of rectangular plates

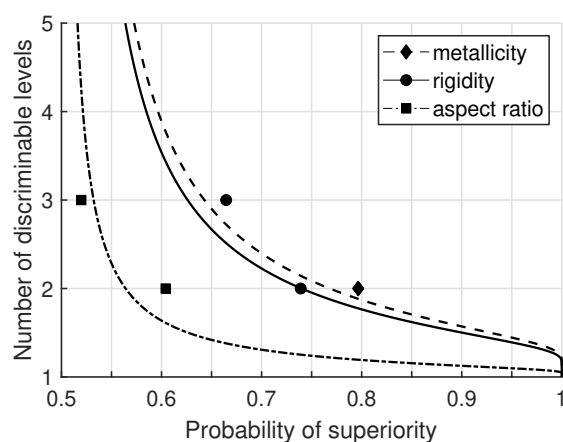


Figure 5.19.: The number of discriminable levels of the three parameters (metallicity, material, aspect ratio), plotted against the probability of superiority P_s . Markers show P_s computed from confusion matrices. Lines show estimates based on effect size.

each parameter: the probability of superiority that is achieved for the number of levels that were tested in the experiment (3 rigidities, 3 aspect ratios, 2 metallicities). In case of two levels, P_s is equal to the accuracy. In case of three levels, P_s is computed individually for both adjacent steps (between levels 1 and 2 as well as between levels 2 and 3). The overall P_s is then the average of the individual P_s between adjacent levels. For material and aspect ratio, the confusion matrix can be reduced to two levels, in order to also obtain P_s for the two-level case. This can be done in two ways, by combining either levels 1 and 2, or 2 and 3, which refers to two different decision thresholds. Assuming that participants would choose the best decision threshold, the maximum of both values is selected. The resulting reference values for two and three discriminable levels are marked in Fig. 5.19.

5.3.5. Discussion and conclusions

Overall, the results show that a combined absolute identification of the three physical parameters metallicity (2 levels), rigidity (3 levels), and aspect ratio (3 levels) was demanding for the participants of the experiment. While overall accuracy was quite high for metallicity and rigidity identification (0.80 and 0.55, respectively), it was only slightly better than chance for aspect ratio (0.36). While higher damping had a negative effect on metallicity and rigidity perception, no significant effect of damping

on aspect ratio perception was found, which may be attributed to the anyway low accuracy in the latter case.

Perceptually, the three parameters were not independent from each other. Participants could hardly discriminate between glass and aluminum ($Acc=0.57$) while plastic/gold and wood/brass discrimination was excellent ($Acc=0.91$). This may be attributed to the high base frequency of glass and aluminum plates, which leads to a small number of audible partials. This is not only due to frequency itself, but even more due to the strong radiation damping for all partials above the critical frequency of the plate. Metallicity is perceptually expressed through (a) damping of some lower partials due to thermoelasticity, and (b) the amount of frequency-dependent damping due to viscoelasticity. (a) is almost completely masked by the small amount of damping that is included, even in the condition with weak damping. (b) affects only higher modes beyond the critical frequency.

A similar, but smaller effect was found for aspect ratio. For bar-shaped plates, it was harder to discriminate between metallicities ($Acc=0.76$) than for compact and longish plates ($Acc=0.82$). This is due to the low modal density for bars compared to plates, making it difficult to judge the damping of higher modes.

Overall accuracy for aspect ratio discrimination was incredibly low, independent of metallicity. However, it was still better for plastic/gold and wood/brass ($Acc=0.36$) than for glass/aluminum ($Acc=0.32$). The latter case was not even significantly different from chance performance. This means that for glass and aluminum, participants were not able to discriminate between different aspect ratios. Following the same argumentation as above, the low number of audible partials makes it impossible to judge the modal density and thus the aspect ratio.

The slightly better performance of trained listeners for rigidity classification could be attributed to the fact that musicians are generally trained in the perception of pitch, i.e., the perceptual parameter that is most closely connected to rigidity.

Within this experiment, the parameters could only take a fixed amount of discrete levels (2 for metallicity, 3 for rigidity and aspect ratio). Those were chosen on the basis of a rough guess of our capabilities in perception. The extreme differences in accuracy for the three parameters show that for an actual sonification, different parameter ranges and/or level segmentations are necessary. These

can be estimated on the basis of Fig. 5.19 and the required probability of superiority P_s . Taking the measured values of P_s for each parameter as reference points, it seems that a computation via Cohen's d already provides a good estimate ($\pm 10\%$), even though the gathered data is only ordinal.

For rigidity, a decimation to two levels within the same parameter range would lead to around 75% accuracy, which seems just enough for sonification. Note that this refers only to absolute judgments without reference. For aspect ratio, even two levels would only lead to 60% accuracy, which surely is insufficient for sonification. In experiment 1, however, we achieved similar accuracy for aspect ratio and length, while length in experiment 1 affected the sound similar to the rigidity in experiment 2. While (1) used aspect ratios between 1.42 and 3.9 (1.5 octaves), in (2) it ranged from 2 to 8 (2 octaves). We therefore assume that a reduced parameter range for aspect ratio (decreased upper limit), in conjunction with a decimation to two levels would lead to sufficient accuracy for sonification. The same applies for rigidity, where the upper limit (glass/aluminum) seems to be high, and a decreased upper parameter limit would improve accuracy.

We assume that two levels may be discriminable for each of the three parameters, in a combined identification task. This leads to a theoretical maximum information capacity of 3 bit or 8 distinct items. Interestingly, this is about the same as was achieved in similar experiments with only one single parameter (Miller 1956). This will be explored in more detail in the next section.

It must be noted that we investigated only absolute identification without reference, similar to absolute pitch perception. From this perspective, even a lower information capacity would seem acceptable for sonification, as most of the information is usually conveyed relatively, as a parameter change over time. Absolute identification of the extreme levels in each parameter within a multi-dimensional sonification is supposed to serve as guidance to facilitate the correct identification of relative parameter changes.

In many practical applications of percussion, we listen for cavities within the examined object. Physicians use percussion to examine and identify cavities inside the human body. Based on the experiences from the above experiment, future research on the information capacity of cavities might be promising.

5.4. The information capacity of multidimensional auditory displays

The results of the previously described experiments seem incredibly bad. In fact, however, we do not have any reference yet, to compare them with. It is therefore necessary to transform the results into a domain that facilitates a comparison with other research, with respect to information theory.

Similar to Pollack (1952) we want to find the information capacity that is conveyed by the auditory display (see also Pollack 1953; Pollack and Ficks 1954). This would make it possible to compare our display with magnitude estimations from other sensory channels, as done by Miller 1956. They all share a similar approach. A given physical or perceptual parameter range is split into a number of discrete levels; participants are then asked to absolutely identify unknown stimuli with respect to the given levels. The same experiment is performed with different resolutions, i.e., numbers of discrete levels that partition the parameter range. The perceived amount of information is then plotted against the sent amount of information. With increasing parameter resolution, the amount of perceived information will converge to a maximum which finally represents the information capacity of the given parameter.

Contrary to this straightforward approach, we performed our experiments only at one single partitioning for each parameter — obviously not enough for fitting a curve and finding a maximum. Under some assumptions of signal detection theory, however, we may be able to find a model to predict the results of other parameter partitionings, based on measured effect sizes, and reformulating the problem of discriminable levels from Sec. 5.2.5.6.

Similar to before, we assume a normal distribution for the parameter estimations on a continuous scale, centered around the true parameter value. In addition, we assume an equal standard deviation for each level. This leads to overlapping standard normal distributions between levels, as visualized for a hypothetical number of 4 levels in Fig. 5.20. We further assume a perfect decision criterion midway between adjacent levels. The tails of the distributions that overlap into the next section defined by the criteria is thus regarded as wrong answer — it describes one level falsely identified as another level. The probability of correct identification is thus the area under the whole probability distribution ($= 1$)

5. Auditory perception and information capacity of rectangular plates

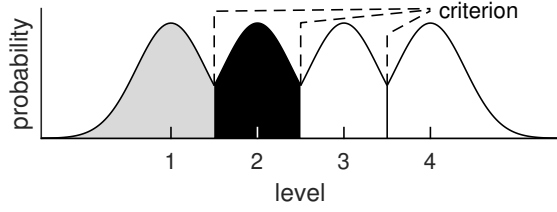


Figure 5.20.: Gaussian model for estimating information capacity.

minus the exceeding tail(s). For the lowest and highest level, this probability P_1 equals

$$P_1 = 1 - \Phi\left(\frac{-|d|}{2}\right), \quad (5.31)$$

as visualized by the gray region in Fig. 5.20. Φ represents the cumulative standard normal distribution. The worst case for the two affected levels, with $d=0$ thus leads to a probability of $P_1 = 0.5$. For the in-between levels (see the black region in Fig. 5.20), the probability P_2 equals

$$P_2 = 1 - 2\Phi\left(\frac{-|d|}{2}\right). \quad (5.32)$$

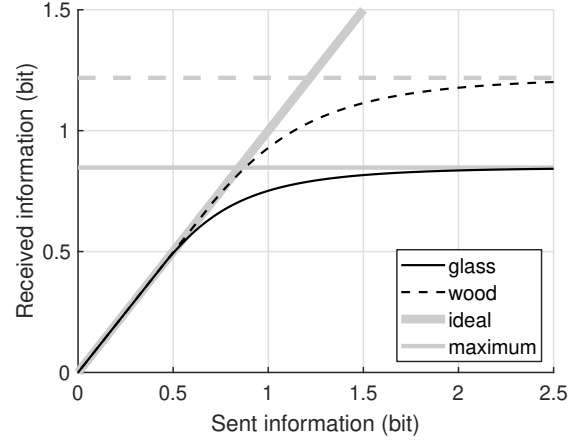
The worst possible probability of correct identification for these levels is thus $P_2 = 0$. Assuming that all levels occur with equal probability (which is the case in our experiments) allows us to simply compute a weighted average of these probabilities, for the given number of levels N , to obtain the final probability of correct identification:

$$PC = \frac{(N-2)P_2 + 2P_1}{N}. \quad (5.33)$$

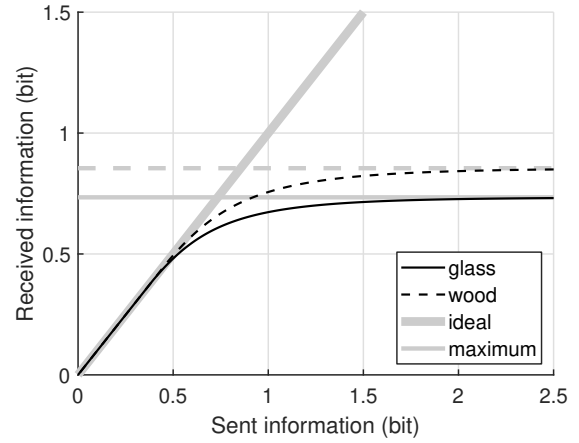
Under the given (admittedly crude) assumptions of the model, for $N = \{2, 3, \dots\}$, the computed values for PC are exact. The number of levels N actually represents the amount of transmitted information $I_{\text{sent}} = \log_2(N)$ bit. The amount of received information I_{rec} is thus:

$$I_{\text{rec}} = \log_2(PC \cdot N) \text{ bit} = I_{\text{sent}} + \log_2(PC) \text{ bit}. \quad (5.34)$$

The plots for the number of discriminable levels of experiment 1 (Fig. 5.9) and experiment 2 (Fig. 5.19) can thus be newly drawn to show the amount of received information vs. the amount of sent information, as done in Fig. 5.21 and 5.22, respectively. For experiment 2, the underlying effect



(a) length



(b) aspect ratio

Figure 5.21.: Received vs. sent information for the two dimensions of experiment 1.

sizes are only approximate. The actual measurement points marked in the plot, however, suggest that the estimated curves are quite close to reality.

For the two-dimensional sonification in experiment 1, we achieve a maximum information capacity of about 1.3 bit for the length parameter, in case of a wooden plate. For the same wooden plate, the information capacity of the aspect ratio parameter is about 0.9 bit. The combined information capacity is then the sum of the two, minus the error that comes from confusing the two parameter dimensions (in contrast to confusing levels within one dimension). The total information capacity of a 2D auditory display with dimensions A and B and confusion accuracy Acc_{AB} is thus

$$I_{\text{rec}} = I_{\text{rec},A} + I_{\text{rec},B} + \log_2(Acc_{AB}) \text{ bit}. \quad (5.35)$$

In case of wood, we achieve a combined information

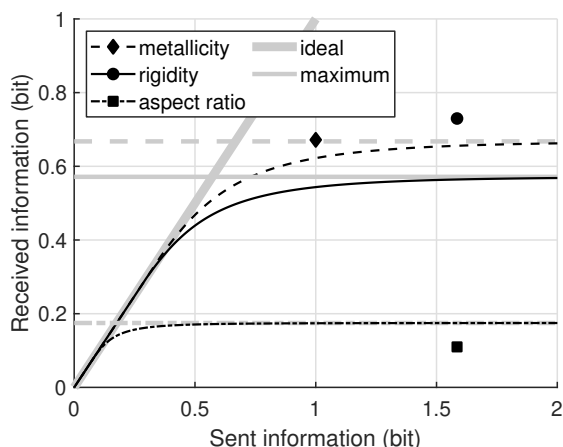


Figure 5.22.: Received vs. sent information for the three dimensions of experiment 2.

capacity of $1.3 \text{ bit} + 0.9 \text{ bit} + \log_2(0.81) \text{ bit} = 1.9 \text{ bit}$.

For experiment 2, the parameter confusion is already contained in the estimate effect size. The total information capacity is thus just the sum of the individual dimensions: $0.7 \text{ bit} + 0.6 \text{ bit} + 0.2 \text{ bit} = 1.5 \text{ bit}$.

How good is that in comparison to the literature? Miller (1956) already showed that the information capacity of a single parameter, no matter from which sensory modality, usually lies between 5 and 9 levels or between 2 bit and 3 bit. This incredibly low capacity is not due to a low perceptual resolution, but rather due to the low capacity of our short-term memory. Pollack (1952) showed that the information capacity for pitch perception is about 2.5 bit, irrespective of the actual frequency range: even if we can perfectly identify and discriminate 5 high pitches and 5 low pitches in two different experiments, a combination of the two frequency ranges again yields only 5 identifiable levels. The same applies for loudness, with an information capacity of 2.3 bit. (Pollack 1952)

It seems logical that a combined two-dimensional auditory display using pitch and loudness would yield the sum of both information capacities, as the two dimensions seem to be almost completely independent. Pollack (1953) showed that we are not living in such a perfect world. The combined number of identifiable combinations is not $5 \times 5 = 25$ but more close to $5 + 5 = 10$; he measured a combined information capacity of 3.1 bit. Taking this into account, our results aren't looking that bad at all. We didn't even reach the 2.5 bit of pitch identification (one dimension) with our 2D

and 3D auditory display, but that was not expected anyway, due to our assumption in the first place: we explicitly chose less salient but more plausible parameter dimensions, knowing that this will reduce the information capacity of the auditory display.

The logical consequence of all this is to circumvent the hard limit of one-dimensional information capacity by adding more and more dimensions to the auditory display. Pollack and Ficks (1954) already took this idea to the extreme by presenting an 8-dimensional auditory display that used 8 independent sound parameters of 2 levels or 1 bit resolution each. By that method, listeners were capable of perceiving almost all of the 8 bit sent: about 7 bit were received.

Our approach to use two or three sound parameters of low resolution was already the right choice, but not yet enough. It must additionally be noted that our physically-inspired parameter dimensions partly interfere with each other, so that participants were possibly forced to distinctly remember every single parameter combination. That may be why our multidimensional auditory displays are still within the range of single parameters such as, e.g., the saltiness of water (1.9 bit, see Miller 1956).

Bibliography

- Aurenhammer, Michael (2021). „Informationsgehalt rechteckiger Platten“. Bachelor's thesis. University of Music und Performing Arts, Graz, Austria.
- Chaigne, Antoine and Christophe Lambourg (Apr. 2001). "Time-domain simulation of damped impacted plates. I. Theory and experiments". In: *The Journal of the Acoustical Society of America* 109.4, pp. 1422–1432. DOI: 10.1121/1.1354200.
- Cremer, Lothar, M. Heckl, and B. A. T. Petersson (2005). *Structure-borne sound: structural vibrations and sound radiation at audio frequencies*. 3rd ed. Springer. ISBN: 978-3-540-22696-3.
- Czuka, Martin (2021). "Sound Synthesis and Acoustic Characterization of Rectangular Plates". Master's thesis. University of Music and Performing Arts, Graz, Austria.
- Czuka, Martin, Marian Weger und Robert Hödrich (2021). „Klangsynthese und akustische Erkennung rechteckiger Platten“. In: DAGA - Jahrestagung für Akustik. Vienna, Austria.
- Fantini, DA and NF Viemeister (1987). "Discrimination of frequency ratios". In: *Auditory processing of complex sounds*, pp. 47–56.

5. Auditory perception and information capacity of rectangular plates

- Groß-Vogt, Katharina et al. (Apr. 5, 2021). "Peripheral Sonification by Means of Virtual Room Acoustics". In: *Computer Music Journal* 44.1, pp. 71–88. DOI: 10.1162/comj_a_00553.
- Howell, David C. (2011). *Confidence Intervals on Effect Size*. Lecture notes. University of Vermont. URL: <https://www.uvm.edu/~statdhtx/methods8/Supplements/MISC/Confidence%20Intervals%20on%20Effect%20Size.pdf> (visited on 04/06/2022).
- International Electrotechnical Commission (2014). *IEC 61260-1:2014 - Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications*.
- Krotkov, Eric (1995). "Robotic perception of material." In: *IJCAI*, pp. 88–95.
- Kuhn, H. W. (Mar. 1955). "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1, pp. 83–97. DOI: 10.1002/nav.3800020109.
- Leeuw, Joshua R. de (Mar. 2015). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser". In: *Behavior Research Methods* 47.1, pp. 1–12. DOI: 10.3758/s13428-014-0458-y.
- Lutfi, Robert A. and Ching-Ju Liu (Aug. 2007). "Individual differences in source identification from synthesized impact sounds". In: *The Journal of the Acoustical Society of America* 122.2, pp. 1017–1028. DOI: 10.1121/1.2751269.
- Lutfi, Robert A. and Christophe N. J. Stoelinga (Jan. 2010). "Sensory constraints on auditory identification of the material and geometric properties of struck bars". In: *The Journal of the Acoustical Society of America* 127.1, pp. 350–360. DOI: 10.1121/1.3263606.
- McIntyre, M.E. and J. Woodhouse (June 1988). "On measuring the elastic and damping constants of orthotropic sheet materials". In: *Acta Metallurgica* 36.6, pp. 1397–1416. DOI: 10.1016/0001-6160(88)90209-X.
- Miller, George A. (Mar. 1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological Review* 63.2, pp. 81–97. DOI: 10.1037/h0043158.
- Neri, Julian and Philippe Depalle (2018). "Fast Partial Tracking of Audio with Real-Time Capability Through Linear Programming." In: *International Conference on Digital Audio Effects (DAFx)*. Aveiro, Portugal, pp. 326–333.
- Pollack, Irwin (1952). "The Information of Elementary Auditory Displays". In: *The Journal of the Acoustical Society of America* 24.6, pp. 745–749.
- (July 1953). "The Information of Elementary Auditory Displays. II". In: *The Journal of the Acoustical Society of America* 25.4, pp. 765–769. DOI: 10.1121/1.1907173.
- Pollack, Irwin and Lawrence Ficks (Mar. 1954). "Information of Elementary Multidimensional Auditory Displays". In: *The Journal of the Acoustical Society of America* 26.2, pp. 155–158. DOI: 10.1121/1.1907300.
- Rossing, Thomas D., ed. (2014). *Springer handbook of acoustics*. 2nd ed. Springer. DOI: 10.1007/978-1-4939-0755-7.
- Ruscio, John (2008). "A probability-based measure of effect size: Robustness to base rates and other factors." In: *Psychological Methods* 13.1, pp. 19–30. DOI: 10.1037/1082-989X.13.1.19.
- Schroeder, Manfred R. (1976). "Machine Processing of Acoustic Signals: What Machines Can Do Better than Organisms (and Vice Versa)." In: *Workshop on Recognition of Complex Acoustic Signals*. Berlin, Germany: Abakon.
- Stoelinga, Christophe N. J. and Robert A. Lutfi (Nov. 2011). "Discrimination of the spectral density of multitone complexes". In: *The Journal of the Acoustical Society of America* 130.5, pp. 2882–2890. DOI: 10.1121/1.3647302.
- Thurlow, W. R. and S. Bernstein (Apr. 1957). "Simultaneous Two-Tone Pitch Discrimination". In: *The Journal of the Acoustical Society of America* 29.4, pp. 515–519. DOI: 10.1121/1.1908946.
- Warburton, G. B. (June 1954). "The Vibration of Rectangular Plates". In: *Proceedings of the Institution of Mechanical Engineers* 168.1, pp. 371–384. DOI: 10.1243/PIME_PROC_1954_168_040_02.
- Weger, Marian et al. (2022). "The information capacity of plausible auditory augmentations: percussion of rectangular plates." In: *Interactive Sonification Workshop (ISon)*. Delmenhorst, Germany.
- Wildes, Richard P. and Whitman A. Richards (1988). "Recovering material properties from sound." In: *Natural computation*, pp. 356–363.

6. “Schrödinger’s box”: an experimental platform for implausible auditory augmentation

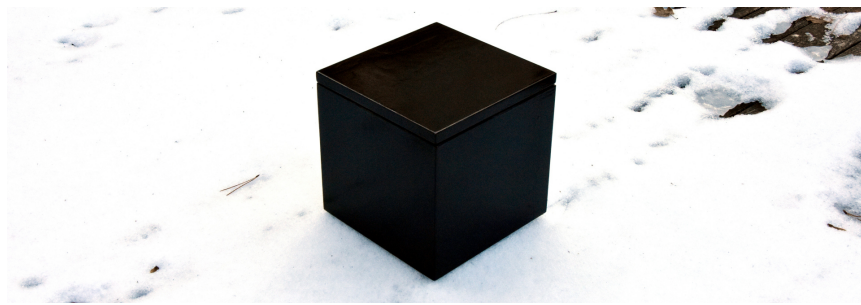


Figure 6.1.: A black box just appeared in the snow. How would it sound, if you struck it with a mallet?



A condensed version of this chapter in the form of an article is provided by Weger, Svoronos-Kanavas, and Höldrich (2022). The sound library and some parts of the software implementation for this project were created by Iason Svoronos-Kanavas during his Erasmus internship under my supervision, based on my original ideas.

We have still no clue about how far plausible auditory feedback can get from the original, and at what point implausible auditory feedback begins. For AltAR/table, we already chose a realistic sound model that is at least physically plausible. For exploring the limits of plausibility, such a physical model is too restrictive. We therefore propose a different platform that is capable of producing even absurd sounds.

Imagine a physical object that doesn’t give any visual indication on its material, its internal structure or contents, its meaning, or its purpose. Something similar to a monolith from Arthur C. Clarke’s “Space Odyssey” novel series or their cinema adaptations. Just like a black box. In fact, it is a black box. “Schrödinger’s box” is depicted in Fig 6.1. How would you expect it to sound, if you struck it with a mallet, as in Fig. 6.2? What kind of sound would be plausible, and what kind of sound would be implausible? That is the question.

From visual inspection, only very few information can be inferred: solid (outer) material, glossy black color, cuboid shape, and a specific size. The most

valuable information is that of solid material. This implies that the object will produce some kind of impact sound in response to the mallet. While size and shape influence pitch as well as intervals between partials, this information unfortunately does not restrict absolute pitch if there are no assumptions on material or content. Also the black color is not providing any relevant information at all, but only occludes everything that is hidden behind, including the internal structure of the object. Is it hollow or solid? Are there any loose parts inside, such as other rigid objects, grains, or fluids? Are there some machinery or electric components inside? Does it contain alien life forms or will it just explode?

For the design of Schrödinger’s box, we followed some very basic rules. According to Norman (2013, pp. 10–13), well-designed products have affordances to guide users intuitively to their proper use. For that reason, we tried to hide every signifier that might reveal information. Furthermore, we applied easily discriminable colors to our internal graphical user interface (GUI) for intuitive debugging and calibration, but chose a uniform black for the physical object itself. In addition, its surface is so glossy that it is even impossible to scratch it. These three aspects intend to focus the user’s attention to the black mallet that is lying besides and thus evoke the irresistible desire to strike the box, and—

finally—resolve the mystery simply by listening to the resulting sound.

The performed action refers to a “hit and listen” approach for object identification (Krotkov 1995). In Sec. 2.2 we already examined how physical parameters are perceived based on such impact sounds. While we are able to extract a lot of useful information, many ambiguities show up. In addition, auditory feedback rarely comes alone. It is usually accompanied by feedback from other sensory modalities such as vision or touch. These information channels are not only processed in parallel, but interact and interfere with each other. However, if interaction is only performed by using a mallet, then the haptic feedback does not reveal more information than that of solid material. Visual feedback anyway doesn’t reveal much information.

Human sound source identification shows that there is much space for alternative auditory feedback, i.e., auditory feedback that is different from the physical truth, but that would suffice our imperfect auditory perception for being perceived as plausible with respect to the performed action. There must be a large variety of alternative auditory feedbacks that are plausible with respect to the little visual and haptic feedback that is returned through the interaction loop.

Schrödinger’s box is a case study for exploring the plausibility of auditory feedback for unknown sounding objects. In the next section, we will discuss plausible auditory feedback for such objects in more detail (Sec. 6.1). Afterwards, we will describe the apparatus, i.e., Schrödinger’s box hardware platform (Sec. 6.3) as well as the software implementation based on onset detection (Sec. 6.4) and sample playback. A discussion of the results follows in Sec. 6.6.

6.1. The plausibility of auditory feedback

The plausibility of sound has already been discussed in Sec. 1.7. What can we thus expect from Schrödinger’s box? As the other senses suggest a solid object, we assume that opposing auditory feedback without a distinct onset yields implausibility. A fluid sound, for example, can hardly lead to a plausible impression—especially keeping in mind that the performed action is almost perfectly identified from sound alone (Lemaitre and Heller 2012). For basic low-level plausibility, the feedback

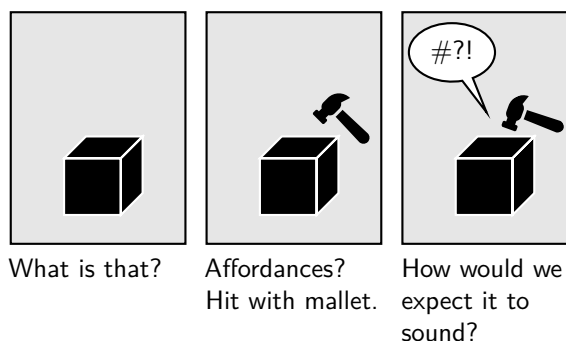


Figure 6.2.: The main research question: what makes auditory feedback plausible?

from different sensory channels (audition, vision, and haptics) should maintain a certain degree of congruency.

A long resonating decay or reverberation tail might be plausible for a hollow box. A squeaking sound of a rubber duck might be plausible for a box made of rubber. The sound of cracking glass succeeding a primary impact might plausibly indicate that the sidewalls are made of thin glass. Rattling or chattering is assumed to be plausible in case of partially loose parts within the box. And even sounds of machinery or living creatures might be plausible under certain circumstances. For example, if the object is bigger than an average cat, it might be plausible if it responded “miao” to a mallet impact (in addition to the sound of a hollow box). We hypothesize, however, that the high complexity of explanation (cat locked inside the box) reduces the plausibility of the sound. If the box roared like a tiger, it is assumed to be implausible, as it is physically impossible to squeeze such an animal in this little box.

What about synthetic sounds? How plausible can they be? For synthetic sounds that mimic a physical system, most literature comes to the conclusion that we cannot discriminate between real and synthetic, even with very simplified modal synthesis models (e.g., Lutfi et al. 2005; Traer et al. 2019). These models are widely used in psychoacoustic experiments, and we feel free to argue that they will be as plausible as their physical archetype.

For synthetic sounds that don’t mimic any physical system, we can expect two possible ways of perception. First, the sound may be identified as artificial, with the result that the only plausible explanation is that of hidden electronics inside the object. Under the assumption that everything can be faked, however, a person will presumably accept just anything

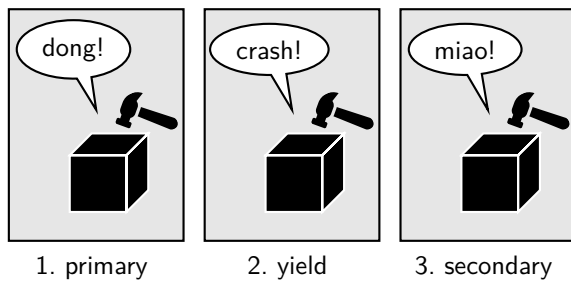


Figure 6.3.: The three sound layers created by Schrödinger's box.

as similarly plausible, and the question of plausibility loses its meaning.

Second, the listener may not be fully aware of the sound being synthetic. The listener might then perceive it as plausible, if it is a cartoonification of a plausible sound. Grimshaw (2009) argues towards the existence of a dip in the plausibility function, similar to the uncanny valley (Mori et al. 2012), for synthetic stimuli that are very close to the true stimulus (see also Wages et al. 2004 and Sec. 1.7). A cartoonification may therefore be even more plausible than a simulation close to reality. This effect, however, has not yet been found outside the visual domain, and even there, it strongly depends on the given context. Cartoonifications help to communicate the specific information that they were designed for, e.g., in movies, video games, or sonifications (Rocchesso et al. 2003). We suppose, however, that if regarded in isolation, i.e., when users are deliberately asked to judge a certain sound, a cartoonification might be recognized as synthetic and thus implausible.

6.2. A taxonomy of sounds for a black box

There already exist various concepts of parameter spaces or timbre spaces of sounds. Assuming that plausible auditory feedback requires a plausible physical explanation, the parameter space itself must be based on physical parameters, too. Such a taxonomy of physical sounds was proposed by Monache et al. (2010). Within this taxonomy, all interactions between solids are based on three low-level models: fracture, impact, and friction. Higher-level processes such as crushing, bouncing, or squeaking, are constructed from these elements. In the same way, musical instruments can be categorized based

on the type of excitation (e.g., bowed or struck) and the type of resonator (e.g., plate or air column), as was done by Lakatos (2000). If every possible sound, irrespective of its physical feasibility, needs to be covered by the timbre space, more abstract sound parameters (e.g., spectral centroid or spectral flux) or their perceptual counterparts (e.g., pitch or roughness) are required. This approach was investigated by Lakatos (2000) and McAdams (2019). Others, e.g., Gounaropoulos et al. (2006), combined perceptual parameters such as brightness, warmth, and harshness with physical properties (e.g., metallic, plucked, etc.).

Given the “hit and listen” approach depicted in Fig. 6.2, we are able to reduce the necessary timbre space of Schrödinger's box to impacts or at least sounds that exhibit a distinct transient. As discussed in the previous section, sounds that are far beyond this category are unlikely to create a plausible multisensory percept. The sounds resulting from mallet impact to a solid black box can be sorted into three categories or layers. These refer to different causal relationships between the user's action and the system's sonic response, as visualized in Fig. 6.3.

Layer 1 corresponds to the primary sound in direct response to the impact. This basic sound depends on impact force as well as on the spatial location of the excitation. In particular, a stronger impact leads to a louder sound with boosted high frequencies. The excitation position defines the distribution of energy into the individual vibrational modes, depending on their respective shapes. Due to geometric patterns of nodal lines (Chladni patterns), this usually leads to some kind of position-dependent comb filtering. The sound may exhibit additional components due to (partially) loose components that may lead to additional impacts and thus rattling or chattering.

Layer 2 includes those sounds that are emitted if the impact force exceeds a certain object-specific threshold. It describes the yield sound that occurs due to material fatigue, leading to crushing/crumpling sounds or even fracture/cracks or complete shattering into loose components. For fluids, it may also refer to the mallet indentation into a fluid in contrast to hitting its surface.

Layer 3 finally refers to secondary sounds that emerge from sub-processes concerning the inner structure or contents. These may be purely mechanical, driven by energy from the impact, but also machinery with its own source of energy, or even result from a living being such a cat locked inside

6. “Schrödinger’s box”: an experimental platform for implausible auditory augmentation

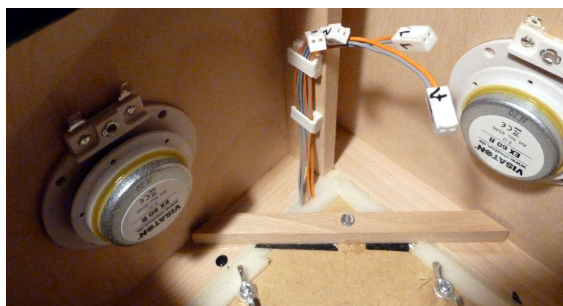


Figure 6.4.: The empty box inside, showing the structure-borne exciters.

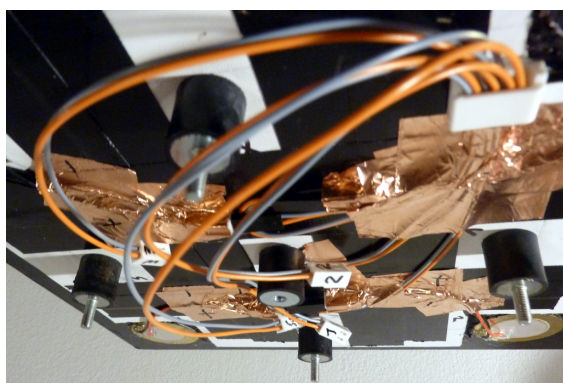


Figure 6.5.: The top plate holding the contact microphones, together with its rubber feet that attach it to the rest of the box.

the box.

For the purpose of future experiments, a sound library has been created specifically for exploring the limits of plausibility with Schrödinger’s box. It utilizes the previously discussed features and intends to cover the whole timbre space ranging from completely plausible or alternative auditory feedbacks up to definitely implausible ones.

6.3. Hardware platform

The black box itself is based on a 23 cm hollow wooden cube. Its basic structure is glued together from strips of beech wood, covered with 4 mm beech plywood on the four vertical sides. These sides act as bending wave loudspeakers, each driven by a Visaton EX 60 S structure-borne exciter, attached from the inside by using 3M VHB tape (see Fig. 6.4). On the top, the cube is first covered by a 3 mm medium-density fiberboard (MDF) plate which then carries the 16 mm MDF top plate via 5 rubber

dampers. Four piezo disks of 3 cm diameter are glued to its bottom near to the corners, acting as contact microphones (see Fig. 6.5). The top plate is reinforced by a smaller 3 mm MDF plate that is glued to its bottom side and exhibits cutouts for the piezos. The dampers in combination with the relatively high thickness and weight of the reinforced top plate minimize acoustic feedback between exciters and contact microphones. The sandwich construction of the top plate further leads to high damping and thus minimizes the original auditory feedback if struck with a mallet. It is assumed that nobody would think of knocking the box on its side.

The gap between main structure and top plate is concealed by strips of solid spruce, so that only about 1 mm is left to maintain decoupling. The whole box is painted in shiny black, so that the wooden construction is not visible anymore, and thus no visual cues of material are given. The glossy surface further ensures that no significant sound can be created by scratching. Under this aspect, even plausible hand interaction might be possible (tapping, knocking), provided that onset detection is sensitive to fingers.

Except contact microphones and exciters, all necessary electronics are mounted on a thin plate that is inserted diagonally into the box from the bottom, with the possibility to remove it for servicing work. Figure 6.6 shows the block diagram of the main components. The basis comprises a BeagleBone¹ Black (Rev. C) microcomputer that is running the Bela software². It is extended by the CTAG FACE cape for analog multi-channel audio and the Bela cape for additional analog and digital sensors (McPherson and Zappi 2015; Langer and Manzke 2018). The four contact microphones are connected to the audio inputs of the CTAG cape through buffer preamps (2× Schatten Design MicroPre 2). The exciters are connected to the audio outputs via class-D amplifiers based on the TPA3116D2 chip. In addition, an ADXL330 3-axis accelerometer is connected to the analog inputs of the Bela cape. Wireless communication and remote development is provided through a USB WLAN dongle.

The three main components (BeagleBone and capes, preamps, amps) are shielded individually from each other, from the WLAN antenna, as well as from outside electromagnetic fields.

The whole system is battery-powered. As the BeagleBone introduces noise to other components

¹BeagleBone: <https://beagleboard.org/bone>

²Bela: <https://bela.io/>

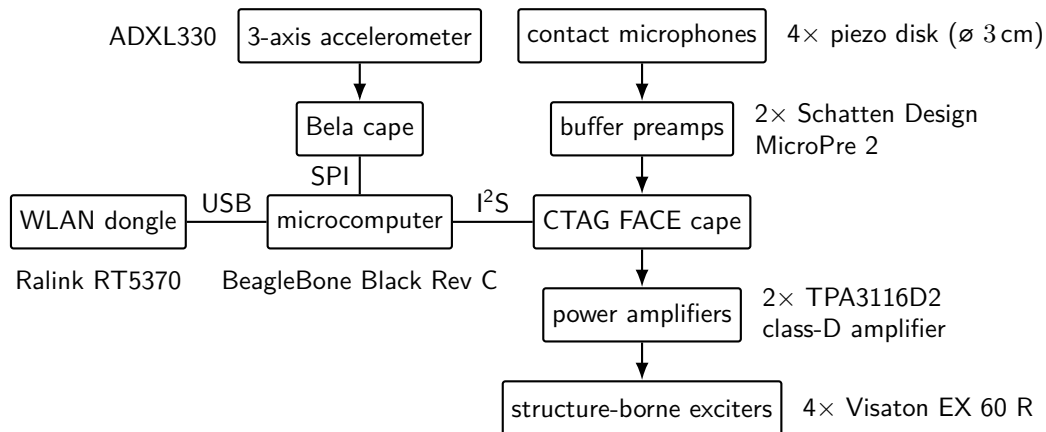


Figure 6.6.: Block diagram of the hardware signal flow of Schrödinger's box. Arrows indicate analog connections.

in case of a common power supply, it is powered by a separate 5 V USB powerbank (EMOS Alpha Slim 10 Ah). This also powers a relay to switch the supply voltage for preamps and amps that is delivered by a Shanqiu FX 5-12 Mini UPS powerbank with 10 Ah capacity (rated at 3.7 V) and separate outputs for the 5 V and 9 V that are required by preamps and amps, respectively. These supply voltages are further filtered by a Yoyo ZGP Noise Blocker for preamps and by two analog low-pass filters (-50.5 dB @ 4 kHz) for amps, respectively. As the peak power of amps and exciters largely exceeds the rated power of the used battery, two 10 mF buffer capacitors are used. If fully charged, the batteries last for several hours of continuous use. The electronics that are placed within Schrödinger's box are shown in Fig. 6.7.

The exciters are equalized by using an inverse filter. Their impulse response (IR) was obtained by the exponential sine sweep (ESS) method (Farina 2000); the signal processing was done in Matlab, based on the Pd implementation by Vetter and Rosario (2011). Measurements were carried out with a pair of Behringer ECM8000 measurement microphones from about 20 cm distance to the box. The sweep was played to all outputs in synchrony; the energetically averaged magnitude spectra of both microphones form the overall magnitude response in Fig. 6.8. Based on the smoothed inverse of the measured magnitude spectrum, a minimum phase filter of 256 samples length was created (Smith 2010, pp. 297–303). The group delay of the filter is as low as 0.15 ms at its maximum near the magnitude notch around 400 Hz. The filter IR is applied in real time by partitioned convolution.

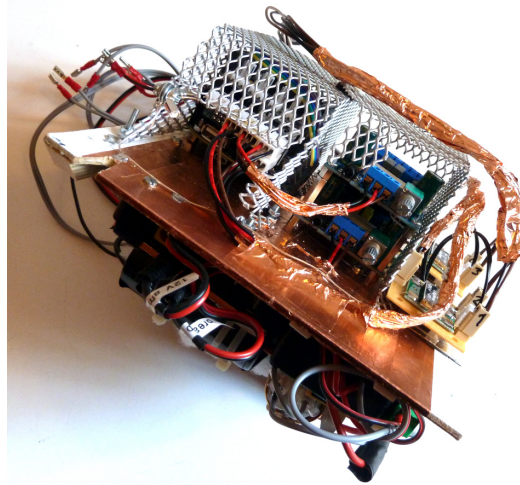


Figure 6.7.: "Schrödinger's guts": the electronics within Schrödinger's box. The components for power supply are mounted on the bottom.

All digital signal processing is implemented in SuperCollider. It consists of two main components: onset detection (see Sec. 6.4) and sample playback (Sec. 6.5). The Bela IDE has been deactivated for saving computing power.

6.4. Real-time onset detection revisited

The augmented object is designed to play a sound sample through the loudspeakers (i.e., structure-borne exciters), as soon as an impact is detected by the contact microphones. All this needs to take

6. “Schrödinger’s box”: an experimental platform for implausible auditory augmentation

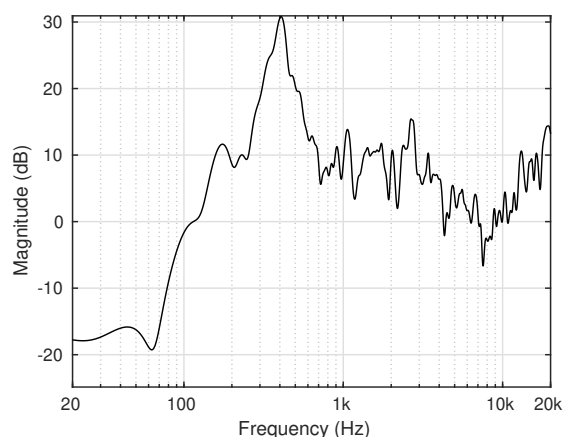


Figure 6.8.: Measured magnitude spectrum of the structure-borne exciters of Schrödinger’s box, with arbitrary normalization.

place as fast as possible, so that the augmented auditory feedback merges to a single sound event with the original auditory feedback.

A delay between the auditory and haptic feedback can be recognized by human listeners if it exceeds 10 ms (Jack et al. 2016). At this latency, a maximum jitter of ± 1 ms is accepted; a larger jitter may be compensated by lower latency (Jack et al. 2016). The total round-trip latency of the system, including onset detection, must therefore strictly stay below this threshold. In addition, the augmented auditory feedback needs to fuse with the original auditory feedback. Our own informal listening test suggests that a latency of less than 10 ms is required for that task. For open-canal hearing aids, a latency below 5 ms is generally recommended in order to avoid audible interferences (Herbig and Chalupper 2010). We therefore target an overall round-trip latency that is at least below 5 ms. Note that the temporal acuity of the auditory system is only between 1 or 2 ms (Green 1971).

6.4.1. Onset time: predict the future and undo the past

Reliable detection of onsets takes time. We are entering the dilemma of causality: an onset can only be detected after it has already occurred. And even worse: according to Dixon (2006), the perceptual onset time occurs already at a level that is between 6 dB and 15 dB below the peak of the onset, depending on the sound material. Optimally, the sample playback would therefore start at this perceptual

onset time—at the latest. Assuming that onsets in our case are always short impacts, with silence everywhere else, we can detect in time domain, if the amplitude reaches a certain threshold. Usually, this approach requires some sort of pre-processing to filter out low-frequency noise and to automatically adapt the threshold to the overall signal level and noise floor. This time-based approach is able to detect onsets almost instantly. However, it may not be very reliable. Due to noise or amplitude modulation, peaks may be falsely identified as onsets, or true onsets might be missed in case that they are masked by some louder stationary low-frequency sound.

That is why most current onset detection algorithms work in frequency domain, based on the short-time Fourier transform (STFT) (Dixon 2006). While these methods are generally robust, an onset can only be detected at the end of the current FFT frame containing the onset. This implies a sample-rate reduction of the onset detection, leading to a theoretical jitter of detection time between 1 sample (in case the onset appears in the last sample of the frame) and the length of one frame (in case the onset appears right at the beginning). In case of an FFT size of 256 samples at 48 kHz sampling rate and non-overlapping rectangular windows, this induces up to 5 ms of jitter. In practice, overlapping Hann windows are used. Assuming that onsets are only detected if they are within the central 50% of the window, between 1 ms and 4 ms of latency are introduced. Frequency-based onset detection algorithms therefore may be robust in detecting onsets, but bear the disadvantage of large temporal jitter, way above our threshold of hearing (see, e.g., Jack et al. 2016 for detection thresholds).

An approach to overcome the jitter was proposed by Turchet (2018) who used time-based onset detection (TBOD) to retrieve the correct time for any onset that is detected by frequency-based onset detection (FBOD). In this case, however, the jitter is compensated by even more additional delay, which is unacceptable in our case. We therefore propose to already trigger sample playback if an onset is detected by TBOD, and then either confirm if FBOD also detects it (hit or true positive), or cancel it if FBOD disagrees (false alarm). Such cancellation is barely noticed since it occurs already within the first 5 ms of playback. FBOD may also trigger playback itself if TBOD overlooked a true onset (miss), or agree with TBOD that there was no onset within the last frame (correct rejection). A block diagram of this approach is shown in Fig. 6.9. Hit and false

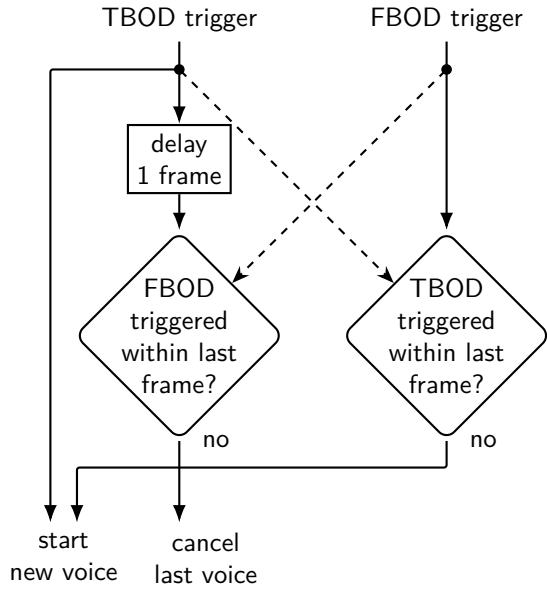


Figure 6.9.: Block diagram of the merging strategy of TBOD and FBOD. The delay equals the FFT size (5.3 ms for 256 samples FFT size at 48 kHz sampling rate).

alarm do not require any additional action by FBOD, as TBOD did alright. In case of false alarm, sound playback must be stopped instantly. In case of miss, FBOD can either trigger a new onset, or decide that it is now anyway too late and simply feel ashamed.

As FBOD is anyway always too late with its detection, no additional delay is introduced for jitter compensation in case of miss. For 256 samples FFT frames with 75% overlapping Hann windows, the onset detection time of this approach theoretically jitters between 0 ms (hit) and 4 ms (miss, onset appears at 75% of FFT frame). Assuming a round-trip latency of 2 ms, this sums up to a latency of 4 ms with ± 2 ms of jitter. While Jack et al. (2016) reported that a random latency of 10 ± 3 ms feels as bad as a constant 20 ms latency, we assume that our high theoretical jitter is compensated by the small constant latency below 5 ms. This assumption, however, has not been tested experimentally.

TBOD is inspired by the time-domain algorithm proposed by Turchet (2018), but tuned to lower latency. It is depicted in Fig. 6.10. Under the assumption that an onset by nature must exhibit a broad spectrum, the input signal first passes a 2nd-order high-pass filter at 4 kHz to remove low-frequency noise. Then, the energy of the signal is approximated by the Teager-Kaiser energy operator

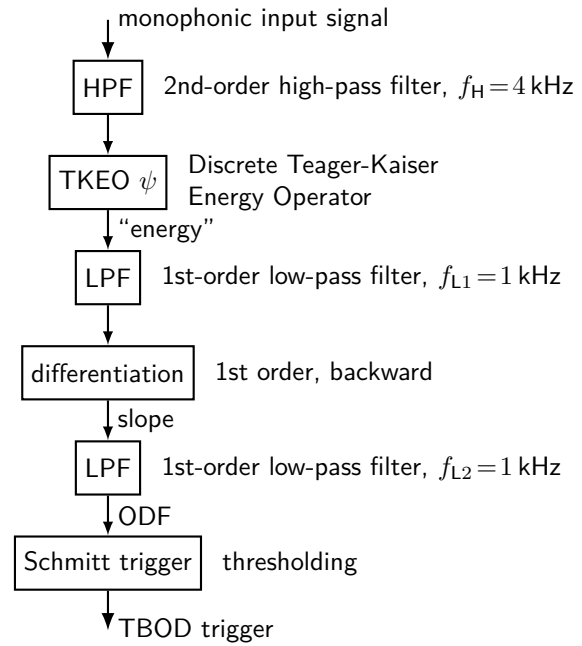


Figure 6.10.: Block diagram of time-based onset detection (TBOD).

(TKEO) ψ (Kaiser 1990; Kahrs 2001),

$$\psi(x[n]) \approx x^2[n-1] - x[n]x[n-2], \quad (6.1)$$

which provides a significant improvement for time-based and frequency-based onset detection, even by just adding it at the beginning of the signal chain (Istvanek et al. 2020). The energy is smoothed by a 1st-order low-pass filter at 1 kHz cutoff frequency. The slope of this smoothed energy is computed by simple 1st-order backward differentiation, to make detection agnostic to the absolute sound level. Another smoothing stage through a 1st-order low-pass filter at 1 kHz yields the onset detection function (ODF). The cutoff frequencies have been selected as a compromise between false alarm rate and latency. Onsets are finally detected by simple thresholding of the ODF by using a Schmitt trigger: a new onset is triggered if its high threshold is exceeded; the next onset can only occur if the signal has fallen below its low threshold. In addition, re-triggering is prevented during a refractory period of 50 ms between successive onsets. With the above settings, onsets are detected within approximately 16 samples.

FBOD is based on the algorithm described by Stowell and Plumbley (2007) and its implementation for SuperCollider in form of the `Onsets UGen`. A block diagram is shown in Fig. 6.11. The input signal is transformed to frequency domain via FFT

6. “Schrödinger’s box”: an experimental platform for implausible auditory augmentation

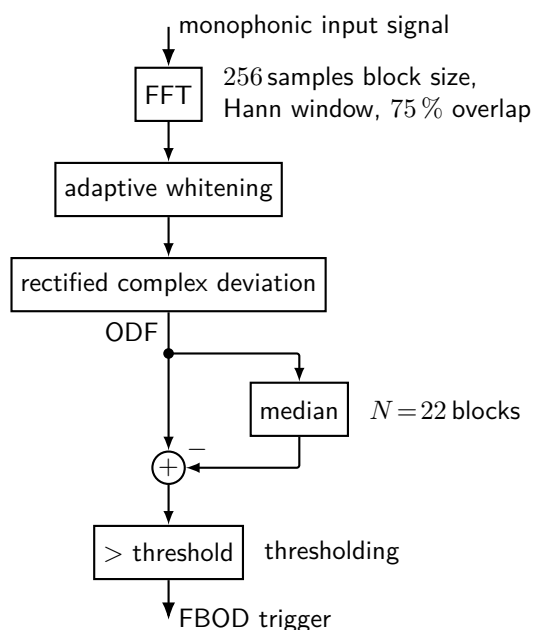


Figure 6.11.: Block diagram of frequency-based onset detection (FBOD).

(256 samples block size, Hann window, 75% overlap), then passes a pre-processing step for adaptive whitening. As ODF, the rectified complex deviation (RCD) is computed. A median filter over 22 blocks (equals 29 ms at the given block size and overlap) retrieves the general trend of the ODF which is then subtracted. Onsets are finally detected by simple thresholding. The next onset can only be triggered after a refractory period of 40 blocks (53 ms).

6.4.2. Impact force and spatial location: extract features before they emerge

For triggering sound samples with plausible sound parameters, we need information on impact force as well as its spatial location on the physical object. The extraction of both features is based on the same measurements of the individual microphone channels as the detection of onset time in Sec. 6.4.1. What we are actually interested in, is the amplitude of the input signal: separately for all input channels, in order to roughly estimate the spatial location of the onset via level differences; averaged over all channels, for controlling the amplitude of the sample playback.

For each channel, the raw input signal is first pre-conditioned through a 1st-order high-pass filter at

1.5 kHz, followed by taking the absolute value. The actual amplitude, i.e., the absolute maximum of the signal envelope, is only reached after some time (here about 0.9 ms after detecting the onset). A straightforward approach would therefore be to wait with sample playback for 0.9 ms and then set the amplitude to the absolute maximum since detection time. As we cannot afford waiting, sample playback is already started at detection time, and the amplitude is constantly updated (at audio rate) through the running maximum during the first 0.9 ms, and then held constant by repeating the last value.

As the measured amplitude is much higher if the impact occurs directly at a microphone position, the raw velocities of all channels are averaged in order to counteract this effect. The average value is then mapped to absolute playback gain of the sound sample (between 0 and 1) through a linear mapping function that is tuned by ear.

For sound localization, only relative level differences are necessary. These are already sufficiently existing at onset detection time to obtain a rough estimate of the spatial location of the onset. Keep in mind that the onset is already detected after about 16 samples after the true onset, and amplitude differences at this time mainly occur due to time differences between the individual channels. The energy ratio between the top/left and top/right microphone, as well as between the bottom/left and bottom/right microphone are measured. A minimum and maximum ratio is defined (about 0.5 and 2, respectively), and the values are then mapped from the exponential range to a linear range between -1 and 1 which represent normalized coordinates on the top plate of the cube. Results from both measurements (top and bottom) are averaged to obtain the Cartesian coordinate in x -direction. The same procedure is performed with the energy ratios between top and bottom in y -direction.

6.5. Sample playback

The sample playback engine of Schrödinger’s box implements the three sound layers that were introduced in Sec. 6.2. While the primary sound is entirely processed in audio rate on the SuperCollider server, so that its playback is instantly started at onset detection time, yield and secondary sound are processed through the SuperCollider language which offers more flexibility in starting new processes on demand.

Primary sound. On every new onset, a new sam-

ple is randomly selected without repetition, out of a sample bank that holds multiple versions. The sample is then piped to the next idle voice of the polyphonic voice allocation. All voices are continuously running in idle to be ready at any time. If all slots are busy (in case of long samples), the least recent voice is faded out within the running refractory period to be available for the next onset. The maximum number of voices is defined individually for each sound preset, as the appropriate number of voices depends on the specific sound. To give an example, several impacts on a rigid object such as a glass plate superimpose, whereas a squeaky toy can only produce one squeak at a time.

The continuously updated amplitude is sent to the latest voice through an audio bus. The spatial location which is available already at onset time, is mapped to a pair of filters that shape the sound in a simple but reproducible way, mimicking the position-dependent energy distribution across vibrational modes. The x - and y -position are identically mapped to the frequency of a matched pair of peaking and notch filters. On the main diagonal of the surface, they cancel each other, but if hit off-center, boost and attenuation occur at different frequencies, shaping the sound while to some extent preserving the original signal energy. Impact force is mapped linearly to the overall playback level. In addition, a boost of high frequencies which would naturally occur with increasing impact force is achieved by a level-dependent shelving filter.

Yield sound. The yield sound implements the same random sample selection and force-dependent output level, but leaves out the additional filtering which would not be physically meaningful in this context.

Secondary sound. The secondary sound is triggered at random after two or more onsets. The probability to trigger increases from 0 to 1 with each impact, starting from the second impact after the end of the last secondary sound. The maximum number of necessary impacts for surely triggering the secondary sound is set individually for each sound preset.

6.6. Conclusions and outlook: exploring the limits of plausibility



Demo videos of Schrödinger's box are available in Source 6.1, while the source code is made public

in Source 6.2. With Schrödinger's box we now have exactly the tool that we need for exploring the plausibility of auditory augmentations. Due to its flexible multi-sampling engine that is tuned to its one and only affordance (being struck by a mallet), it is capable of providing plausible as well as implausible auditory feedback. Its total round-trip latency (including A/D and D/A conversion, onset detection, and sample playback) equals about 2.4 ms at a hardware buffer size of 16 samples. It therefore seems to meet the extreme requirements of low-latency signal processing that are needed in order to merge the augmented auditory feedback with the original auditory feedback to a single sound percept. While this assumption needs further systematic examination, informal listening tests suggest a major improvement in comparison to 5 ms round-trip latency.

With regards to the combination of frequency- (FBOD) and time-based (TBOD) onset detection for triggering samples, it was observed that false alarms as well as missed onsets rarely occurred, even with TBOD alone, for the given mallet interaction. In addition, the latency of FBOD was observed to be unacceptable for the given task. FBOD is therefore deactivated in laboratory situations, but kept as an option for applications in noisy environments such as public sound installations.

In the future, several experiments are planned on the basis of Schrödinger's box. First, the fusion of original auditory feedback and augmented auditory feedback to a single sound percept needs to be measured as a function of loudness ratio and latency. Second, the limits of plausibility should be explored via absolute judgments of plausibility in combination with verbalizations similar to the estimation of causal uncertainty (Ballas and Sliwinski 1986; Lemaitre et al. 2010), in order to finally derive an auditory plausibility analysis model similar to PAM (Connell and Keane 2006).

Bibliography

- Ballas, J. A. and M. J. Sliwinski (Nov. 1, 1986). *Causal Uncertainty in the Identification of Environmental Sounds*. ONR-86-1. Fort Belvoir, VA: Defense Technical Information Center. DOI: 10.21236/ADA175228.
- Connell, Louise and Mark T. Keane (Jan. 2, 2006). "A Model of Plausibility". In: *Cognitive Science* 30.1, pp. 95–120. DOI: 10.1207/s15516709cog0000_53.

6. “Schrödinger’s box”: an experimental platform for implausible auditory augmentation

- Dixon, Simon (2006). “Onset Detection Revisited”. In: *International Conference on Digital Audio Effects (DAFx)*. Montréal, Canada, pp. 133–137.
- Farina, Angelo (2000). “Simultaneous measurement of impulse response and distortion with a swept-sine technique”. In: *AES Convention*. Paris, France: Audio Engineering Society.
- Gounaropoulos, Alex et al. (2006). “Synthesising Timbres and Timbre-Changes from Adjectives/Adverbs”. In: *Applications of Evolutionary Computing*. Ed. by Franz Rothlauf et al. Red. by Josef Kittler et al. Vol. 3907. Springer Berlin Heidelberg, pp. 664–675. DOI: 10.1007/11732242_63.
- Green, David M. (1971). “Temporal auditory acuity.” In: *Psychological Review* 78.6, pp. 540–551. DOI: 10.1037/h0031798.
- Grimshaw, Mark (2009). “The audio Uncanny Valley: Sound, fear and the horror game.” In: *Games Computing and Creative Technologies*.
- Herbig, R. and R. Chalupper (2010). “Acceptable processing delay in digital hearing aids.” In: *Hearing Review* 17.1, pp. 28–31.
- Istvanek, Matej et al. (Jan. 4, 2020). “Enhancement of Conventional Beat Tracking System Using Teager–Kaiser Energy Operator”. In: *Applied Sciences* 10.1. DOI: 10.3390/app10010379.
- Jack, Robert H., Tony Stockman, and Andrew P. McPherson (2016). “Effect of latency on performer interaction and subjective quality assessment of a digital musical instrument”. In: *Audio Mostly*. Norrköping, Sweden: ACM Press, pp. 116–123. DOI: 10.1145/2986416.2986428.
- Kahrs, Mark (2001). “Audio Applications of the Teager Energy Operator.” In: *AES Convention*. Espoo, Finland: Audio Engineering Society.
- Kaiser, J.F. (1990). “On a simple algorithm to calculate the ‘energy’ of a signal”. In: *International Conference on Acoustics, Speech, and Signal Processing*. Albuquerque, NM, USA: IEEE, pp. 381–384. DOI: 10.1109/ICASSP.1990.115702.
- Krotkov, Eric (1995). “Robotic perception of material.” In: *IJCAI*, pp. 88–95.
- Lakatos, Stephen (Oct. 2000). “A common perceptual space for harmonic and percussive timbres”. In: *Perception & Psychophysics* 62.7, pp. 1426–1439. DOI: 10.3758/BF03212144.
- Langer, Henrik and Robert Manzke (Apr. 29, 2018). “Embedded Multichannel Linux Audiosystem for Musical Applications”. In: *Journal of the Audio Engineering Society* 66.4, pp. 286–291. DOI: 10.17743/jaes.2018.0022.
- Lemaitre, Guillaume and Laurie M. Heller (Feb. 2012). “Auditory perception of material is fragile while action is strikingly robust”. In: *The Journal of the Acoustical Society of America* 131.2, pp. 1337–1348. DOI: 10.1121/1.3675946.
- Lemaitre, Guillaume et al. (2010). “Listener expertise and sound identification influence the categorization of environmental sounds.” In: *Journal of Experimental Psychology: Applied* 16.1, pp. 16–32. DOI: 10.1037/a0018762.
- Lutfi, Robert A. et al. (July 2005). “Classification and identification of recorded and synthesized impact sounds by practiced listeners, musicians, and nonmusicians”. In: *The Journal of the Acoustical Society of America* 118.1, pp. 393–404. DOI: 10.1121/1.1931867.
- McAdams, Stephen (2019). “The Perceptual Representation of Timbre”. In: *Timbre: Acoustics, Perception, and Cognition*. Ed. by Kai Siedenburg et al. Vol. 69. Cham: Springer International Publishing, pp. 23–57. DOI: 10.1007/978-3-030-14832-4_2.
- McPherson, Andrew P. and Victor Zappi (2015). “An Environment for Submillisecond-Latency Audio and Sensor Processing on BeagleBone Black”. In: *AES Convention*. Warsaw, Poland: Audio Engineering Society.
- Monache, Stefano Delle, Pietro Polotti, and Davide Rocchesso (2010). “A toolkit for explorations in sonic interaction design”. In: *Audio Mostly*. Piteå, Sweden: ACM Press, pp. 1–7. DOI: 10.1145/1859799.1859800.
- Mori, Masahiro, Karl MacDorman, and Norri Kageki (June 2012). “The Uncanny Valley”. In: *IEEE Robotics & Automation Magazine* 19.2, pp. 98–100. DOI: 10.1109/MRA.2012.2192811.
- Norman, Donald A. (2013). *The design of everyday things*. Revised and expanded edition. New York, New York: Basic Books. ISBN: 978-0-465-05065-9.
- Rocchesso, Davide, R. Bresin, and M. Fernstrom (Apr. 2003). “Sounding objects”. In: *IEEE Multimedia* 10.2, pp. 42–52. DOI: 10.1109/MMUL.2003.1195160.
- Smith, Julius Orion (2010). *Physical audio signal processing: For virtual musical instruments and audio effects*. W3K publishing. ISBN: 978-0-9745607-2-4.
- Stowell, Dan and Mark Plumbley (2007). “Adaptive whitening for improved real-time audio onset detection.” In: *International Computer Music Conference (ICMC)*, pp. 312–319.

- Traer, James, Maddie Cusimano, and Josh H. McDermott (2019). "A perceptually inspired generative model of rigid-body contact sounds". In: *International Conference on Digital Audio Effects (DAFx)*. Birmingham, UK.
- Turchet, Luca (2018). "Hard real time onset detection for percussive sounds." In: *International Conference on Digital Audio Effects (DAFx)*. Aveiro, Portugal, pp. 349–356.
- Vetter, Katja and Serafino di Rosario (2011). "ExpoChirpToolbox: a Pure Data implementation of ESS impulse response measurement". In: *Pure Data Convention*. Weimar, Germany.
- Wages, Richard, Stefan M. Grünvogel, and Benno Grützmacher (2004). "How Realistic is Realism? Considerations on the Aesthetics of Computer Games". In: *Entertainment Computing (ICEC)*. Ed. by M. Rauterberg. Vol. 3166. Springer Heidelberg, pp. 216–225. DOI: 10.1007/978-3-540-28643-1_28.
- Weger, Marian, Iason Svoronos-Kanavas, and Robert Höldrich (2022). "Schrödinger's box: an artifact to study the limits of plausibility in auditory augmentations." In: *Audio Mostly*. St. Pölten, Austria: ACM.

7. Prospects and limits of auditory augmentations



This chapter is based on the original publication which was created in teamwork by Groß-Vogt, Weger, and Höldrich (2018), as well as on my associated talk at Forum Media Technology 2018. My main contributions to SBE4 include parts of technical planning, preparation, and execution.

While experimenting with auditory augmentations, several questions arose that seem to be crucial for their design:

1. Which kinds of *data* are suitable?
2. What are the *characteristics* of auditory augmentations?
3. Which kinds of *objects* are suitable, and can a room serve as an object in this context?
4. Do we need a new *definition* of auditory augmentation?

These questions cannot be answered by a single person or working group alone, without the help of the research community. In order to shed some light on these research questions, we organized an interdisciplinary workshop with renowned researchers coming from different fields of sonification, design, sound art, composition, and engineering.

7.1. An interdisciplinary workshop on auditory augmentation

To explore the prospects and limits of auditory augmentation, we initiated a practice-oriented prototyping workshop with an interdisciplinary group of researchers from various institutions. This 4th edition of the Science by Ear (SBE) workshop series and parts of its results are also described by Groß-Vogt et al. (2018) and on its website¹. While previous instances focused on certain types of data (e.g., physics data or climate data), this was the first instance to focus on a certain sonification method.

¹SBE4: <https://iem.at/sbe4>

The basic idea of the SBE workshop series is to bring together researchers with different backgrounds to explore a certain problem of sonification. The about 20 invited participants may comprise sound or sonification experts with programming skills, composers and sound artists, or researchers of a certain domain science that is connected to the topic (see Fig. 7.1). The participants are working in interdisciplinary teams in which each member takes one of the following roles: programmer, sound expert, data expert, or moderator. The three- to four-day workshop alternates between prototyping sessions in groups and plenum discussions where the developed prototypical sonifications are presented and discussed with the other participants. Within the prototyping sessions, each team approaches their given task through brainstorming, data listening, verbal sketching, concept development, and programming, in order to present a prototype within limited time.

The previous instances of the SBE series showed that great care is to be taken to carefully balancing the creative setting with well-prepared tasks. The focus on auditory augmentation in SBE4 implied that data and software had to be prepared by the organizers and also understood by the participants. Especially the possibilities and restrictions of the prepared systems had to be explained. For quick prototyping, there were well-prepared code examples in SuperCollider for data processing (indexing, averaging, filtering, mapping, etc.), as well as fully work-

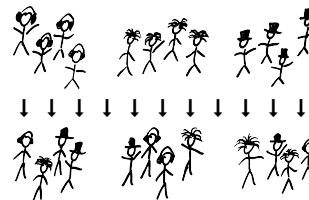


Figure 7.1.: The concept of the Science by Ear workshop series: sonification experts, artists, and domain scientists forming interdisciplinary teams for quick prototyping.

7. Prospects and limits of auditory augmentations

ing examples for each combination between three prepared sonification platforms and three kinds of data.

Concerning the workshop format, SBE4 mastered the problems of too many options and too little preparation, which occurred during the previous instances of the SBE series. The major improvements comprised well prepared datasets and platforms, different complementing datasets and platforms, and a balance between restrictions and freedoms. Altogether, this led to a great success of SBE4.

The 19 participants of SBE4 comprised 11 researchers that can be counted to the sonification community; nevertheless, with varying backgrounds in sciences, arts, and humanities. The remaining 8 included two media and interaction experts, two composers, two sound engineers, a musicologist, and a sociologist. Twelve participants were at post-doc level or above, while seven were pre-doc researchers or students. Unfortunately, only three of the participants were female. During the prototyping sessions, the group sizes varied between three and six, as not all participants were able to attend the whole program.

During the three days of the workshop, each group created one prototype for each of three different hardware platforms. They were free to choose any of the three datasets that allowed to implement their specific ideas.

7.2. Datasets and sonification platforms

7.2.0.1. Datasets

Concerning data, the SBE4 workshop concentrated on in-home electric power consumption. This type of data was selected due to its relevance to almost everyone, and due to its scalability, from the battery level of the cellphone in the pocket to detailed information on the power consumption of all households within a city or country. For the workshop, we prepared three types of data, aiming at different levels of data analysis.

The *personal monitoring* data was based on real-time measurements of 5 individual kitchen appliances: dish washer, coffee machine, water kettle, microwave, and fridge. The measurements were done via wall plugs of the Fibaro intelligent home system (see Fig. 7.2) which sent measured power wireless over the z-wave protocol. The sampling interval was 1 s, with measurements in the range



Figure 7.2.: Wall plug and wireless USB dongle for real-time electric power measurement.

between 0 and 3000 W. The data was received by a computer through an Aeotec Z-Stick USB dongle and by using the free and open-source Domoticz² home automation system. The data was forwarded to the network via OSC, so that it could be received in real time from any computer within the wireless network.

For *personal data exploration*, we prepared the recorded data of one private household, in which the loads of 9 individual appliances have been measured during one year, with a sampling interval of 10 minutes. The measured appliances include water kettle, TV, light and electronics in the living room, fridge, toaster, dehumidifier, dishwasher, and washing machine.

For *big data analysis*, we prepared a large dataset that stems from smart meter data in Ireland, and includes measurements from 12 000 households over a period of 1.5 years, at a sampling interval of 30 min, tagged with additional metadata (Commission for Energy Regulation 2012). From this large dataset, we extracted consistent data of 54 households for each combination of three family structures (single, couple, family), two education levels (secondary vs. tertiary), and two types of housing (apartment vs. detached house).

We pre-processed both offline datasets so that missing data points have been fixed.

7.2.0.2. Sonification platforms

Each type of data could be explored in combination with one of three different platforms for auditory augmentation. While one of these (*sensors*) is a multi-purpose sensor platform that can be freely moved and attached to any physical object, the other two (*table* and *room*) are more restricted in that they specify both the physical object and the technical setup.

²Domoticz home automation system:
<https://www.domoticz.com/>



Figure 7.3.: The BRIX₂ physical computing and sensor platform. Photo: Sebastian Zehe, 2014.⁵

The *sensors* platform is the most versatile of the provided platforms. It basically consists of the BRIX₂ physical computing platform developed by Zehe et al. (2012).³ BRIX₂, as shown in Fig. 7.3, basically consists of an Arduino⁴-compatible microcontroller with on-board sensors that is packed into a Lego-compatible housing. It can be easily extended by additional sensor modules and offers USB and wireless communication. Its simple design in combination with well-documented code examples allow rapid prototyping, perfectly suited for the SBE workshop.

The *table* platform (see Fig. 7.4) basically is an early version of the AltAR/table platform which is described in detail in Sec. 4. It represents an auditory augmentation of a table, in which the auditory appearance of the table can be set to any solid and isotropic material such as metal or glass, by changing the physical properties of the underlying physical sound model. The model parameters (size, shape, material, etc.) can be arbitrarily changed with respect to external data, spatial location of the interaction, and time. The physical interface comprises a wooden plate that is equipped with hidden contact microphones and exciters or additional loudspeakers, and a marker-based optical tracking system to locate the position of any object or hand interacting with the surface. Any sound that emerges from physical interaction with the interface plate is augmented in real time, to allow a plausible change of the perceived materiality while, e.g., knocking or writing on it.

The *room* platform builds upon the IEM Cube, a multi-purpose laboratory, performance, and lecture hall (see Fig. 7.5). First described by Zmolnig et al. (2003), it is equipped with a 24-channel loudspeaker array, arranged on a half sphere for ambisonic sound spatialization. It features a virtual room acoustics

³<https://www.techfak.uni-bielefeld.de/ags/ami/brix2/>

⁴<https://www.arduino.cc/>

⁵Source: <https://opensource.cit-ec.de/projects/brix2/wiki/LiBRIX%E2%82%82>



Figure 7.4.: The table platform in SBE4, as seen from above.



Figure 7.5.: The IEM Cube, serving as the room platform in SBE4. Photo: W. Hummer / KUG.

system that is driven by five microphones mounted on the ceiling. The artificial reverberation can be controlled in real time via OSC. In the context of SBE4, we allowed live sound input from the microphones, but also the use of additional ambient sounds or soundscapes.

7.3. Prototypes

During the three days workshop, each team created one prototype for each of the three different sonification platforms. The dataset could be chosen freely. The nine prototypes that participants came up with are briefly described in the following. A fourth prototyping session on the last day allowed a refinement of certain prototypes.

7.3.1. “Writing resonances”

Primary task: writing and placing objects.

Secondary task: personal data monitoring.

Metaphor: the higher the load, the bigger the table.

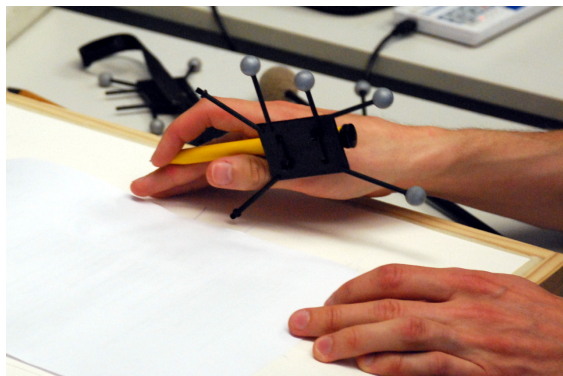


Figure 7.6.: Writing resonances.

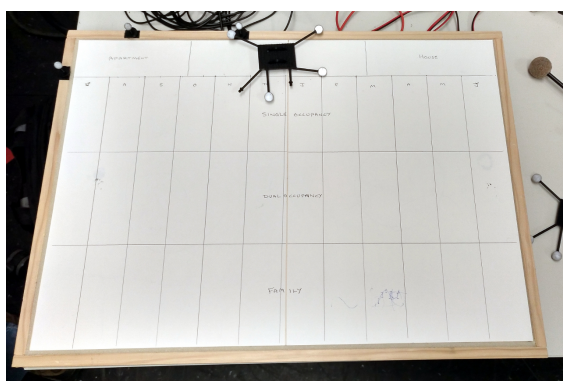


Figure 7.7.: Exploration table.



Writing Resonances (see Fig. 7.6 and video Source 7.1) is a classical application of auditory augmentation, very close to the Wetterreim system, i.e., the augmented computer keyboard by Bovermann et al. (2010). It sonifies the personal monitoring dataset in real time through the table platform. In particular, the overall power consumption of all five appliances is mapped to the size of the model plate, employing the metaphor of a larger table when more power is consumed. The resulting auditory feedback will become deeper in pitch with increasing power consumption. As the table doesn't produce a sound by itself, the data is only conveyed during interaction with the table. This can be writing with a pen, placing or moving objects, and even typing on a mechanical computer keyboard that is placed on the table. While the primary task is writing, information on the power consumption can be accessed any time by knocking on the table.

During the fourth, open session of the workshop, this prototype has been extended so that additional information concerning the different appliances in-

dividual power consumption is conveyed by a modulation of individual partials of the sound model.

7.3.2. "Exploration table"

Primary task: data exploration.

Secondary task: —

Metaphor: —

	apartment				apt. + house				house			
single	1	2	3	4	5	6	7	8	9	10	11	12
couple	1	2	3	4	5	6	7	8	9	10	11	12
family	1	2	3	4	5	6	7	8	9	10	11	12
	month →											

Figure 7.8.: Spatial partitioning of the exploration table interface.

The Exploration Table prototype is shown in Fig. 7.7 and video Source 7.2. It aimed at the exploration of the big dataset on the energy consumption of Irish households, by using the table platform as an interactive interface. The interface allowed to explore the monthly averages of households (from January to December) depending on two dimensions of metadata: type of housing (apartment, detached house) and family structure (single person, couple, or family). Two of the data dimensions are mapped to the two-dimensional surface of the interface plate, resulting in a grid of 12 months and 3 family structures (see Fig. 7.8 for a schematic representation of the spatial partitioning). In order to access also the 3rd data dimension (type of housing), an additional row was added at the top of the plate, divided into three cells representing apartments, houses, or both together. This data dimension is navigated by placing an optical marker that is recognized by the tracking system. The data is explored by interaction with the plate through a tracked felt mallet. The auditory augmentation of the table is changed as a whole, depending on the spatial position of mallet and marker. The three-dimensional data space can thus be explored by using both hands—one to set the housing dimension, and the other to set the time and family structure dimension as well as to induce sound into the physical model of the plate. The data that is conveyed by the sound is then the absolute energy consumption which is mapped to the pitch of the table: higher load is represented by higher pitch.



7.3.3. “Sonic floor plan”

Primary task: recreation

Secondary task: data monitoring

Metaphor: open window represents high load or energy waste.

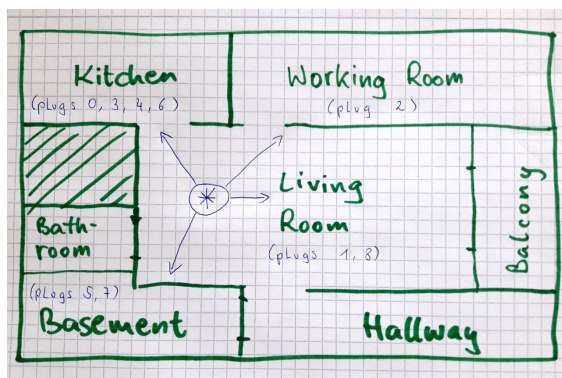


Figure 7.9.: Sonic floor plan.

The Sonic Floor Plan is an auditory display of individual appliances' load within a household. It is based on the personal monitoring data and the room sonification platform. The team assumed a theoretical floor plan of the apartment, in which every appliance is signified at its true spatial location (see Fig. 7.9). The auditory display is supposed to be installed in the living room, preferably in the center of the apartment. From this perspective, every appliance is defined by its direction from the listening position, and its instantaneous load. To prevent an overflow of unnecessary information, only the appliance with the (at the moment) highest load is sonified by a sound that occurs periodically after a specified time, as well as every time a person enters the room. The sound is spatialized so that it is perceived from the direction towards the appliance, with its loudness mapped to its load level. The sound content itself consists of the environmental soundscape from outside the building, captured by a microphone. The system employs the metaphor of a window being opened in the direction of the appliance, thus metaphorically releasing precious energy. Due to the display of only the largest appliance, even small standby consumption is made perceptible in case that no major energy consumer is active.

7.3.4. “Smart kettle”

Primary task: tea/coffee preparation

Secondary task: data monitoring

Metaphor: the kettle rewards or punishes the user



Figure 7.10.: Smart kettle.

The Smart Kettle uses the real-time data of the energy consumption of a kettle, in order to display the amount of wasted energy when boiling water, expressed by the amount of surplus water that is left in the kettle after preparing a tea, coffee, etc. In the prototype, as shown in Fig. 7.10 and video Source 7.3, the kettle is equipped with a BRIX₂ microcontroller to sense spatial rotation and thus detect the process of pouring water out of it. The use of an additional temperature sensor is contemplated to a future sonification of the temperature of remaining water inside the kettle.

When the kettle is activated, the consumption data is accumulated, and a sound of boiling water is played to signify its state of heating even before the start of the actual physical sound emitted by the active kettle. As soon as the kettle has finished and gets deactivated, the total energy that has been consumed for heating the water gets assessed. If a user then grabs the kettle to pour the boiling water out, a sound of pouring water is played (in addition to any possible actual sound of pouring water). When putting it back on its platform, it announces the amount of wasted energy with an auditory icon. The auditory icon signifies the amount of wasted energy as well as a binary information on low waste (rewarding sound) and high waste (punishing sound).

7.3.5. “Standby door”

Primary task: opening the door

Secondary task: data exploration

Metaphor: —

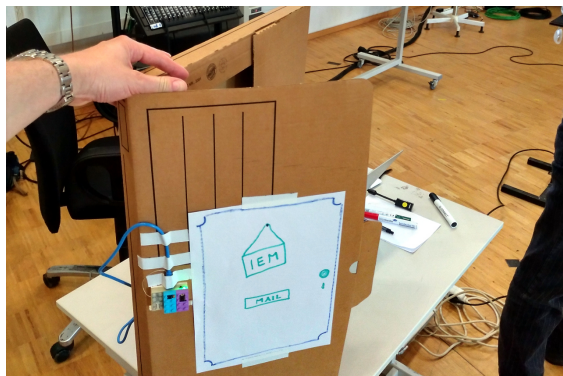


Figure 7.11.: Standby door.

The Standby Door is a design study of an augmented door that displays energy consumption of electric appliances within the household when it is opened. The prototype is based on a cardboard mock-up of a door (see Fig. 7.11), but is intended to be applied to the entrance door of an apartment or house. If the door is opened, the consumed energy of each individual appliance within the last days is displayed. Several sound mappings were developed (see also video Source 7.4).



In a first approach, each individual appliance is represented by a synthetic sound event that is constructed with a one-to-many mapping: the higher the consumption, the higher the pitch, the longer the duration, the louder the sound. The individual sounds are played one after another at a constant rhythm, while the speed with which the door is opened controls the playback speed. This refers to the assumption that a user who opens the door in a fast movement is in a hurry and thus has less time for the sonification. A problem that came up with this sonification is that a certain appliance can only be identified within the sonification by its position in sequential playback.

The second sonification approach aimed at a better differentiation between individual appliances. The sound was therefore based on the spoken names of the individual appliances. The speech was generated by a speech synthesizer. As with the previous sonification, the absolute amount of energy was mapped to pitch and loudness. In addition, the actual mapping was tuned individually to the different

appliances according to their minimum and maximum level of consumption. The team members also discussed the use of more expressive speech synthesizer or recordings from actors, in order to display the data through different levels of emotion (e.g., soft vs. aggressive speech) or voicing (e.g., unvoiced/whispering vs. voiced speech), instead of or in addition to the rather abstract sound parameters pitch and loudness.

7.3.6. “3D gestural mouse”

Primary task: data exploration

Secondary task: —

Metaphor: —



Figure 7.12.: 3D gestural mouse.

The 3D Gestural Mouse prototype uses the BRIX₂ microcontroller board itself as a controller to navigate through the big dataset of Irish households (see Fig. 7.12 and video Source 7.5). Attached buttons allow switching between data dimensions such as household types, education level, and types of housing. The weekly pattern of energy consumption is mapped to the frequency and amplitude of a sine-tone, and played back in a loop. The orientation of the controller let the user change sonification parameters such as base frequency and modulation depth. If the controller gets shaken, a higher speed of shaking leads to an increase in playback speed. During the plenum discussions, the idea arose to change playback speed in discrete steps by rotating gestures clockwise (increase) and counter-clockwise (decrease).



7.3.7. “Hob assistant”

Primary task: cooking

Secondary task: data monitoring (position)

Metaphor: cooking pot snaps in at correct position

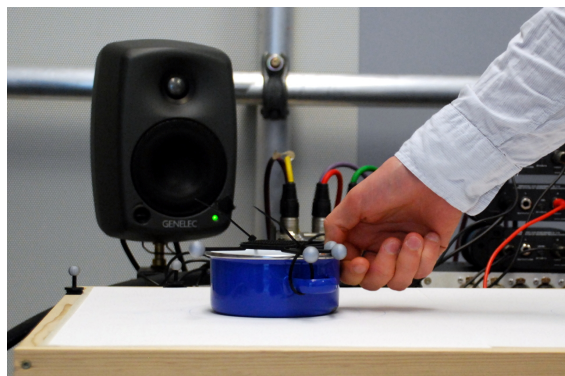


Figure 7.13.: Hob assistant.

The Hob Assistant prototype attempts to sonify how well a cooking pot is centered on a hob of a stove, and how well the size of the selected hob matches the size of the pot. A photo is shown in Fig. 7.13 and video Source 7.6. Building on the table platform, it actually doesn't use any of the provided datasets, but instead tries to assist people to economize their energy use in everyday life. The auditory augmentation of the plate (thought of as the top of a stove in this context) is controlled only by the spatial position of the cooking pot on the hypothetical stove, with hobs signified by circles drawn on its surface.

If the cooking pot is placed far away from any hob, then the auditory augmentation is not active at all. The level of the augmented auditory feedback increases with decreasing distance to the nearest hob. In addition, a feedback-delay (echo) is added to the sound. It exhibits a long delay time for large distance, while the delay time gets shorter when approaching the next hob with the cooking pot. If the pot is placed in the sweet spot of a hob, the very short delay time induces the metaphor that the pot mechanically snaps in at the correct position.

The overall sound of the augmented table is set to short decay and dense overtones, mimicking a kind of plastic material. However, in the vicinity of a hob that fits the size of the cooking pot, the decay time and amplitude of a single partial are raised to create a satisfying sound with distinct pitch to signify the correct hob for the given cooking pot.

7.3.8. “Kitchen sounds”

Primary task: recreation

Secondary task: data monitoring

Metaphor: room grows bigger with higher load

The Kitchen Sounds prototype uses the room platform to sonify real-time data from the individual kitchen appliances. The 5 appliances are represented by metaphoric sounds (samples from a sound library) that are spatialized to different directions in the horizontal plane. The sounds are equally spaced to cover the whole circle (72° separation between the 5 appliances). A one-to-many mapping is applied so that the individual load of each appliance is mapped to the width of its spatial image as well as to the density and duration of the grains of a granular synthesis based on the individual sounds. A higher load leads to denser and longer grains, while a low load leads to very sparse grains of short duration. In addition, the artificial reverberation system is used to display the overall energy consumption via the reverberation time: the metaphor that is employed is that of a larger room in case of a larger load. The one-to-many mapping strategy was chosen to overcome the little information capacity of the individual sound parameters, so that a change in load gets more salient.

7.3.9. “Interleave”

Primary task: data exploration

Secondary task: —

Metaphor: —

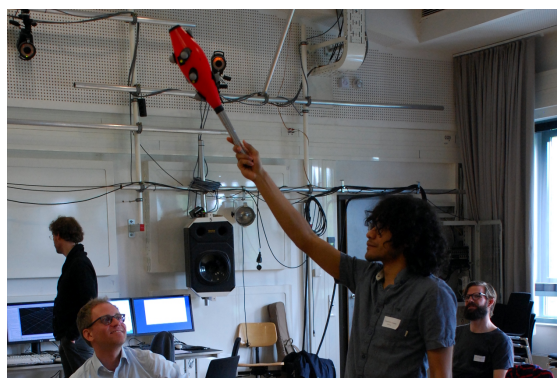


Figure 7.14.: Interleave.

Interleave is an interactive sonification of the big dataset of Irish households, based on the room platform. Its goal is to make the periodicity of the data audible. This may be daily, weekly, monthly,

and yearly patterns. An assumed application is data analysis in a power supply company, e.g., to inspect the load on a certain node. The playback speed can be adjusted to zoom into different cycles, i.e., temporal levels.

The data is displayed as audification: the raw time series of consumption data are simply taken as audio signals that are played back without further processing, in order to make as few assumptions on the data as possible. In addition, the load value is mapped to a transposition via frequency shifting of the signal, as well as to its spatialization. Higher load is mapped to the front, while lower load gets spatialized towards the back, along a ring of loudspeakers in the horizontal plane.

The sonification thus allows a comparison between two different groups of households. The two different subsets of the data (e.g., secondary vs. tertiary income households) are played back together, one spatialized to the left half-circle, and the other to the right half-circle.

The differences in load patterns are especially audible in the diverging rhythmic patterns. For instance, the differences in the average structure of the day (e.g., division of sleeping, working, and after-work time) can be perceived between the two levels of education, based on the daily and weekly patterns.

A refinement of the sonification allows the use of a juggling club that is tracked by the optical tracking system to serve as a controller for setting sonification parameters (see Fig. 7.14 and video Source 7.7). The height of the juggling club sets the playback speed (and thus also controls pitch) between 30 minutes and one month per second.



7.4. Discussion

The prototypes that have been developed within the fourth edition of the Science by Ear workshop cover great amounts of the broad field of auditory augmentation, and also exceeded it in some cases. Together with the plenary discussions of the prototypes revealed individual peculiarities of the individual platforms and datasets, and also of auditory augmentation as such.

7.4.1. On the peculiarity of sound in augmented reality

In all of the developed prototypes, additional sound is embedded into the physical environment of the

user, in order to convey information. This can be regarded as some sort of augmented reality: some digital data is materialized in the form of sound in the physical world. One might argue that any sound conveys data, at least some information on the sound itself. So, is then listening to the radio already augmented reality? In order to be consistent with the visual domain, we must negate this question. It is generally accepted that an overlay of a video on top of a visually perceived scene is not augmented reality, except that there is a direct connection between them (Azuma 1997). The auditory equivalent would be sound that is added to an auditory scene. The radio, without any direct connection to the original auditory-visual scene, therefore cannot be regarded as augmented reality. All the developed prototypes, however, have something that connects them to the physical world: the sound is linked either directly to a user action, or metaphorically to the state of a physical object.

While the above example suggests that both domains (auditory and visual) can be treated more or less similarly, they are in fact radically different. A visualization can be perceived at a glance. A lengthy sonification, however, cannot. In other words: vision takes space, sound takes time. Vision is absolute, sound is relative. These challenges have already been discussed by Kramer et al. (1999) and still bear mysteries. Our visual perception has high resolution in space, but low resolution in time (e.g., cinema works with 24 Hz visual sampling rate). Our auditory perception has high resolution in time, but low resolution in space. Auditory augmentation therefore works best if it is coupled to something that takes time: action. Actions can be neither performed nor perceived at a glance. They take time, just like the sounds that accompany them. Objects that don't act are ignorant to time, and might be better augmented visually. Sonification is for the living. Auditory augmentation is living itself—it cannot be consumed passively, but requires active participation.

7.4.2. So what is auditory augmentation?

A central goal of the SBE4 workshop was to evaluate our working definition of auditory augmentation, and to adjust it, if necessary.

Auditory augmentation is the augmentation of a physical object and/or its sound by sound which conveys additional information.

No participant of the workshop questioned this definition, we would therefore propose it for future applications. However, it seems very broad and covers even systems which the participants unanimously regarded as “not anymore auditory augmentation”. The Exploration Table fell under this category, as well as Interleave, or the 3D Gestural Mouse. These have one thing in common. Their primary application, the primary task, is data exploration. There is no secondary task. This implies that the augmented object has no other use than data exploration. The object is thus not augmented, but rather specifically created for its own purpose.

In case of Interleave or the 3D Gestural Mouse there never was an original purpose. These therefore can't be augmentations.

In case of the Exploration Table, one may argue that it is still a table. Well, indeed it is, but the visual grid, markers, and mallet interaction obviously limits its use. It is practically unusable, and therefore lost its original purpose. If an object's original purpose is changed by the sound, then we might call that an auditory transformation, but not auditory augmentation.

The main task doesn't necessarily be a goal-oriented activity such as writing with the Writing Resonances prototype, or cooking with the Smart Kettle or Hob Assistant; however, one participant noted that “having a concrete task helps to design”. For Sonic Floor Plan, Kitchen Sounds, and Standby Door, the primary task is actually not defined (opening a door is an action that is part of a broader activity). It is more a daily routine or “state of being”, as one participant called it. In case of Sonic Floor Plan or Kitchen Sounds, the task comprises simply being in the living room. However, it is important that this task, whatever it is, is not disturbed by the augmentation. We can therefore phrase a first amendment to the definition of auditory augmentation:

*Auditory augmentation needs a primary task.
This can be anything, but not data exploration.
The task should not be disturbed by the augmentation.*

Concerning the task, there is another aspect that all the valid auditory augmentations of the workshop have in common. The auditory augmentation added a secondary task that is related to the data: monitoring. Furthermore, the monitoring task did not disturb the main task, whatever it was. In some cases, the user might pause its main task to pay more attention to the sonification. If this happens,

the secondary task becomes the main task, and the user might not even just monitor the data, but explore it. For example, with Writing Resonances, the user might interact with the table in different ways (knock, scratch, etc.) to explore different aspects of the data. With the Standby Door, the user could repeatedly open and close the door, to get more precise information on the sonified data. The second amendment thus must be:

Auditory augmentation adds a secondary monitoring or data exploration task.

7.4.3. The relationship between sound, data, and augmented object

Another goal of the SBE4 workshop was to explore the relationships between task, object, and sound, in the context of auditory augmentation. The components of an auditory augmentation are: interaction, object, data, and sound. Within the developed prototypes, these components were interconnected quite differently.

The most basic type of auditory augmentations, in line with the original concept of Bovermann et al. (2010), is structured as depicted in Fig. 7.15. Interaction with the object produces sound, whereas this sound is modulated by the data. This we call *sound-centered* auditory augmentation.

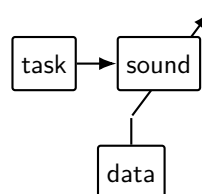


Figure 7.15.: *Sound-centered auditory augmentation. The task produces a sound (auditory feedback) that is modulated by the data.*

If the auditory augmentation employs the metaphor that a physical property of the augmented object is changed (e.g., the size of the object), then it may also be regarded as *object-centered*. In this case, the object itself is metaphorically modulated by the data (see Fig. 7.16). This applies for Writing Resonances and Kitchen Sounds (size of plate or room). In the Sonic Floor Plan prototype, a metaphoric window is opened where valuable energy leaks out, displayed by sound that is coming in.

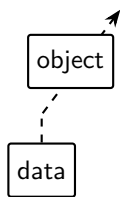


Figure 7.16.: Object-centered auditory augmentation. The object itself is (metaphorically) modulated by the data.

Task-centered auditory augmentations have a specifically linear structure. The task, or more precisely, the actions that comprise it, create data. This data is then sonified. An example is the Smart Kettle, where the data on energy waste is inherent to the task of tea preparation. The same applies for the Hob Assistant, where the sonified positioning data is actually describing the positioning action itself. A block diagram is shown in Fig. 7.17.



Figure 7.17.: Task-centered auditory augmentation. The task produces data that is used to generate sound.

In case of the Standby Door, we observe a *data-centered* approach to auditory augmentation. The sound is created solely by the data. The action (here: opening the door), just defines how the sound is played back. The action modulates the sound (see block diagram in Fig. 7.18).

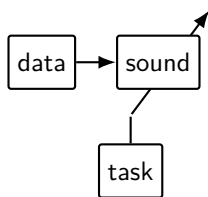


Figure 7.18.: Data-centered auditory augmentation. Data generates sound that is modulated by the task.

The listed approaches to auditory augmentation cannot be seen as strict categories. In fact, some of the prototypes might fit in multiples of these. Figure 7.19 shows an attempt to combine all the possible relationships into one single block diagram. The user action may create sound or data, but also control the sound design of the system. All three

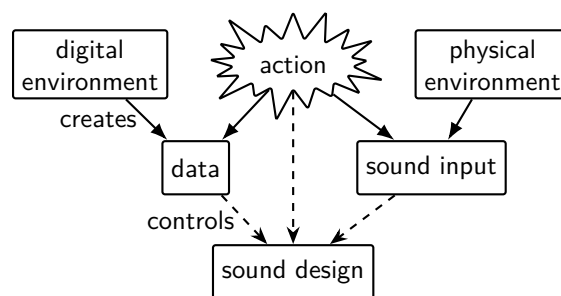


Figure 7.19.: Block diagram of auditory augmentation.

parts (user action, data, and sound input) may control the sound design.

Not all connections are necessary at the same time. The prototypes demonstrated that even very simple relationships might be enough for successful auditory augmentation. However, we assume that many coherent relations between user action, input sound, external data, and sound design make the auditory augmentation more natural and intuitive.

We might summarize the main findings of this section as follows:

Auditory augmentation favors a tight connection between the primary task or action and the resulting sound. This task or action is not necessarily connected to the conveyed data.

Auditory augmentation favors a metaphor which links the sound to the physical state of the object.

In addition, we can conclude on the final research question: can the room be an object for auditory augmentation? Well, of course. The prototypes Kitchen Sounds and Sonic Floor Plan both used the room as an object, and were regarded as unquestionable examples of auditory augmentation by the workshop participants. And also the Knock Knock system by Tünnermann et al. (2013) used the room as an object. All three systems made use of a metaphor concerning the physical properties of the room (room size, open window, and empty room, respectively). It seems that this is indeed a powerful tool for creating plausible and usable auditory augmentations.

7.4.4. Perceptual implications of auditory augmentation

The results of the SBE4 workshop revealed interesting insights on the connection between data and

auditory augmentation. One aspect is standing out, if the prototypes are analyzed with regard to the sonified data. Those prototypes which used the big dataset of Irish households were never classified as auditory augmentations by the participants. Of course, it was already obvious beforehand that this combination would be difficult, but the prototypes clearly showed that no matter what platform (sensors, table, room), the participants completely failed in their goal of creating auditory augmentations with this kind of data. It must therefore be regarded as entirely incompatible with auditory augmentation. This has two reasons. First, the data has such a high dimensionality that the task of navigating within the data takes the entire attention of the user. The data exploration task thus automatically becomes the main task, as happened with Exploration Table, 3D Gestural Mouse, and Interleave. Second, the big dataset is static. While there surely is a time axis, the time is actually frozen. It represents an absolute time instant in history, somehow similar to the one moment of life in the ancient city of Pompeii that has been preserved under the ashes. In other words: the big dataset is non-real-time, it is dead. That's why it doesn't connect well to living action in real time that is strictly required for auditory augmentation.

A borderline case might be the Standby Door. It also displays recorded data from the past. However, the data in this application is directly connected to the user's past actions. This doesn't mean that auditory augmentations strictly require a connection between data and user. The Standby Door might still be a great auditory augmentation, if it displayed completely unrelated non-real-time data. This deficit, however, is compensated by the fact that it neither deteriorates the main task, nor the purpose of the object. Data exploration in this case is added as a secondary task, without disturbing the main task at all.

The problem with data dimensionality and resolution has been thoroughly discussed in the sonification community (e.g., Campo 2007). De Campo's design space map summarizes, which type of sonification might be appropriate for a given number of dimensions and number of data points. It shows clearly that many dimensions and data points are compatible with only one type of sonification: model-based sonification. One might argue that an auditory augmentation that incorporates a physical model *is* model-based sonification.

However, there is one aspect missing: what is going to be perceived by the user. While model-based

sonification is perfectly capable and also highly suitable for the exploration of high-dimensional data such as the big Irish dataset in the workshop, the individual dimensions are then usually not anymore perceived as such. Perception-wise they suffer the loss of detail, the distinct data points get blurred, and form some kind of auditory gestalt that only allows to obtain an overall impression of the data. When it comes to high-dimensional data, sonification is therefore capable of delivering a rough overview at a glance in a way that visualization is not.

The true problem with auditory augmentation therefore isn't the high complexity or dimensionality of the data. All these might be nicely mapped to the model that runs in the background. The problem is the navigation through these dimensions. All three problematic prototypes included some sort of navigation through the data that cannot be done as a secondary task. The Standby Door, however, showed that it is indeed possible to create an interface for data exploration that integrates nicely in the user's physical environment. It is a blended sonification, whereas the other three interfaces are not.

It is therefore erroneous to argue that auditory augmentation needs low-dimensional data. The point is that the auditory augmentation must be monitored or explored in a secondary task, without disturbing the main task, even if the main task is not specified, or just "being" or "behaving". What is actually limited by this constraint is the information capacity: the amount of information that is conveyed by the sonification. And even if the information is encoded in the sound, there are limitations in how much information a user can perceive and process within a secondary task. The difficulty to overcome these problems already resides in the design process of auditory augmentations. One participant raised the issue that designing ambient displays means designing for the background, while in the designer's mind the sound is in the foreground. The designer of auditory augmentations usually has an excitement for sounds. This problem is actually not new and applies to any other field as well: the designer should cultivate a beginner's mind. For the prototypes this was not always easy, as for demonstration purposes, the sounds had to be designed much more prominent, salient, and less calm than an actual application would require. On the other hand, participants noted that "ordinary people" might need more salient sounds than the designers who knew what to listen for. It was therefore

7. Prospects and limits of auditory augmentations

suggested to incorporate some kind of cartoonification, at least for those augmentations that used iconic sounds; otherwise they might “all sound the same” for the typical user.

There is another aspect that came up during the workshop, especially with the Writing Resonances prototype. The relation between physical object and augmented auditory feedback may be perfectly plausible and united in a single multisensory gestalt for a single data point or position in time. However, if the sound changes over time, the object loses its gestalt identity and divides into the true physical object and the sound that is thereby perceived as separate auditory event played from the loudspeakers. The fact that this perceptual effect had not been noticed with the Sonic Floor Plan, might give a hint to its explanation. The table platform, and also the variable room acoustics in the room platform, employ the metaphor of a change in physical parameters of a usually rigid and time-invariant physical object. Especially the table cannot morph into another material or size in our physical world. This contradicts the general laws of physics. The metaphor of the open window in case of the Sonic Floor Plan, however, is physically feasible. We can therefore conclude that auditory augmentations benefit from a metaphor that describes a physically feasible causal relationship between object and sound.

7.4.5. Why auditory augmentation?

All along its existence, sonification has been coping with its right to exist. Entire book sections have been dedicated to this topic (e.g., Pinch and Bijsterveld 2012, pp. 249–270). It therefore seems appropriate and even necessary to also justify auditory augmentation as a very specific and still unacquainted area in the broad field of sonification. This was even one of the main research questions we intended to evaluate. When and why should a designer consider auditory augmentation for a specific product or intention?

The participants of the SBE4 workshop came up with useful scenarios for the pre-defined platforms (sensors, table, room). Most of the developed prototypes reached a promising state already after a three hours prototyping session. One participant summarized: “there are nine prototypes that are really worthwhile considering and working on in the future.” Sure, the strict specifications of the prepared platforms and datasets narrowed down possible paths of design. However, Participants

generally agreed that the hands-on sessions were enriching and provided sufficient time to try out even approaches that bore the risk of not working out in the end.

For example, the team that developed the Writing Resonances prototype asked themselves if the benefit of the auditory augmentation would even be worth the electric energy it consumes itself. With this question in mind, they measured the energy consumption of the table platform in real time, to map it to the sound of the table. Besides the constant load which doesn't bear information if regarded on its own, an audible increase in load was noticed if user interaction and thus augmented auditory feedback was strong. This interaction-dependent load could be attributed to the active loudspeakers which draw more current in case of louder sound. The members of the team concluded that they had built a self-representative system. The auditory augmentation at least needs to refer to the augmented physical object, such as in the Hob Assistant prototype, but not only to the augmentation itself.

That being said, the six valid auditory augmentations that were developed within the SBE4 workshop clearly represent useful applications, and show that auditory augmentation has the power to contribute additional value to everyday interactions, if some simple design principles are maintained.

Bibliography

- Azuma, Ronald T. (Aug. 1997). “A Survey of Augmented Reality”. In: *Presence: Teleoperators and Virtual Environments* 6.4, pp. 355–385. DOI: 10.1162/pres.1997.6.4.355.
- Bovermann, Till, René Tünnermann, and Thomas Hermann (Apr. 2010). “Auditory Augmentation”. In: *International Journal of Ambient Computing and Intelligence* 2.2, pp. 27–41. DOI: 10.4018/jaci.2010040102.
- Campo, Alberto de (2007). “Toward a Data Sonification Design Space Map”. In: *International Conference on Auditory Display (ICAD)*. Montréal, Canada, pp. 342–347.
- Commission for Energy Regulation (2012). *CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]*. 1st ed. Irish Social Science Data Archive. SN: 0012-00.
- Groß-Vogt, Katharina, Marian Weger, and Robert Höldrich (2018). “Exploration of Auditory Augmentation in an Interdisciplinary Prototyping

- Workshop". In: Forum Media Technology. St. Pölten, Austria, pp. 10–16.
- Kramer, Gregory et al. (1999). *Sonification Report: Status of the Field and Research Agenda*. International Community for Auditory Display.
- Pinch, Trevor and Karin Bijsterveld, eds. (2012). *The Oxford handbook of sound studies*. New York, NY: Oxford University Press.
- Tünnermann, René, Jan Hammerschmidt, and Thomas Hermann (2013). "Blended sonification—sonification for casual information interaction." In: *International Conference on Auditory Display (ICAD)*. Lodz, Poland.
- Zehe, Sebastian, Tobias Grosshauser, and Thomas Hermann (Mar. 2012). "BRIX - An easy-to-use modular sensor and actuator prototyping toolkit". In: *International Conference on Pervasive Computing and Communications Workshops*. Lugano, Switzerland: IEEE, pp. 817–822. DOI: 10.1109/PerComW.2012.6197624.
- Zmolnig, Johannes, Winfried Ritsch, and Alois Sontacchi (July 2003). "The IEM-cube." In: *International Conference on Auditory Display (ICAD)*. Boston, MA: Georgia Institute of Technology, pp. 127–130.

8. Case studies of auditory augmentations

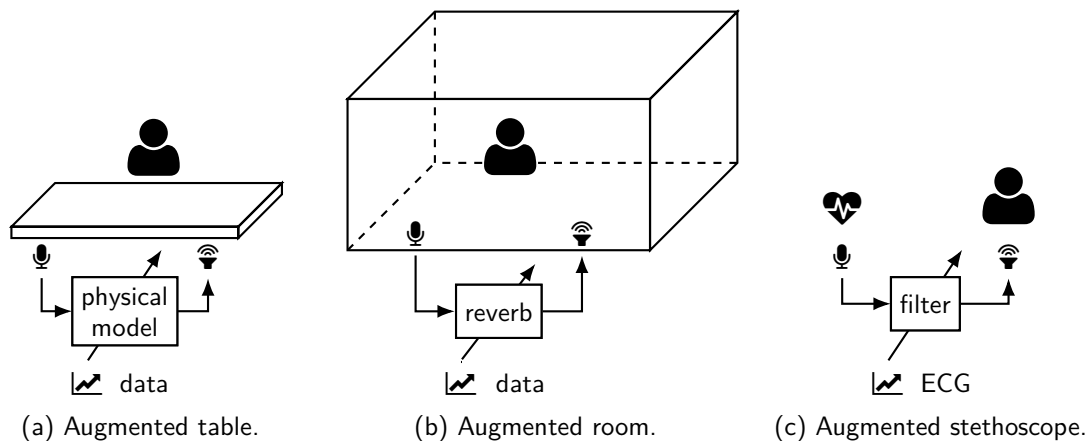


Figure 8.1.: Case studies of auditory augmentations: table, room, and human body.

In the context of this thesis, a number of prototypes and case studies of plausible auditory augmentations have been designed, created, and evaluated. Some of these have already been discussed in the previous chapters. For auditory augmentation of a table, we presented the AltAR/table platform (Ch. 4), with its perceptual evaluation in Sec. 5.2. With Schrödinger’s box (Ch. 6), even implausible auditory augmentations have been explored, even though the actual experiment is still pending at the time of writing. In addition, the 4th Science by Ear workshop (SBE4) yielded a large number of extraordinarily versatile prototypes (Ch. 7), based on AltAR/table, virtual room acoustics, and sensor-equipped domestic appliances.

The possibilities and flavors of the augmented table (Fig. 8.1a), are manifold, and not all of them fit into the scope of this dissertation. AltAR/table actually builds on the experience gathered with a preceding platform: the Mondrian table. Its two prototypical versions, the augmented graphic tablet and the auditory coloring book, are discussed in Sec. 8.1.

The other most promising approach for plausible auditory augmentation is the room (Fig. 8.1b). It has already been studied during SBE4 (Ch. 7), but needs formal evaluation. Therefore, two experiments have been conducted to investigate its

acceptance and comprehensibility (Sec. 8.2) as well as its perceptibility in an ambient scenario (Sec. 8.3).

Finally, a totally different kind of physical object is investigated in Sec. 8.4: the human body. With the aim of facilitating medical diagnosis through auscultation, a stethoscope was augmented by additional information from the electrocardiogram (see Fig. 8.1c).

8.1. “Mondrian table”: display of spatial information by auditory augmentation

This section is based on the following publication: Weger and Höldrich (2019).



In order to explore the potential as well as the constraints of plausible and usable auditory augmentation, we developed an experimentation platform which adheres to the horizontal, even, rigid, rectangular, and stationary surface, as exemplified in Sec. 1.2. The Mondrian table platform was a first prototype of the augmented table that was later pushed to its limits by AltAR/table in Ch. 4. It was intended to offer two different modes of operation. On the one hand, it should enable calm and unobtrusive blended sonification outside the focus of attention while performing daily activities

8. Case studies of auditory augmentations

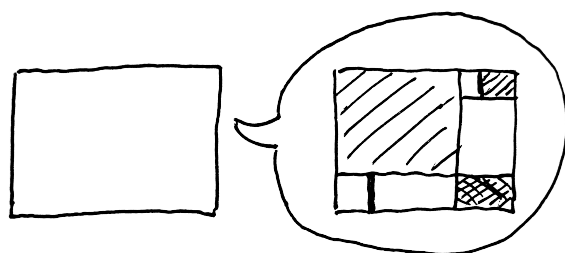


Figure 8.2.: Concept of the Mondrian Table, inspired by Piet Mondrian's "Composition II in Red, Blue, and Yellow" (1930).

that affect it (e.g., writing or positioning of other physical objects). On the other hand, it should serve as an interface for exploratory data analysis through manual interaction in the form of tapping, scratching, etc.

Even on an ordinary dining table, the tabletop can be composed of several elements of different materiality. Such structures can be simulated through a space-dependent auditory augmentation. Our experimentation platform might therefore be perfectly suitable for sonification of a time-invariant configuration of geometric structures, displayed as regions with different perceived physical properties. This relates to a macro surface texture (Müller-Tomfelde and Münch 2001). The underlying data can be a map, a technical drawing, or a painting similar to Piet Mondrian's "Composition" series. We therefore call this experimentation platform the Mondrian Table (see Fig. 8.2).

The Mondrian Table ambiently sonifies geometric structures as plausible and usable changes of the surface material. By placing a sheet of paper on top of it, those structures can be traced back with a pencil and thus transferred into the visual domain.

The data does not directly produce sound, but is mapped to properties of the intermediate physical model (e.g., material category or shape). This model serves as a filter, altering the original sounds emerging from the interaction between user and physical object.

Tünnermann et al. (2013) developed a standard for visualizing the audio- and dataflow of blended sonifications in a simple way. Such a blended sonification diagram usually involves the three factors user, physical environment, and digital environment as main sources of data (D) and/or audio (A). Connections of these sources to the Auditory Display as a sink visualize causal contributions to either a filtered (F) or an added (A) output. While fil-

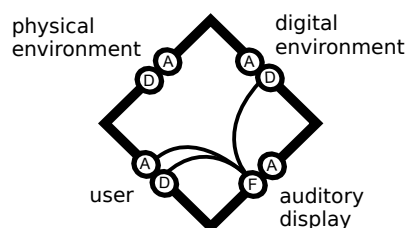


Figure 8.3.: Blended sonification diagram of the Mondrian Table.

tered sonifications are "sonifications that stay very close to the original sound", superimposed "sound samples or synthesized sound fall in the 'added' category" (Tünnermann et al. 2013, p. 4).

The blended sonification diagram of the Mondrian Table (Fig. 8.3) shows that our system produces a filtered sonification which is dependent on audio from the user in order to sonify data from the digital environment. The auditory display additionally depends on properties of the interaction (spatial location) as well as on properties of the physical object (original physical properties). We interpret the physical environment as something that is out of reach and not influenced by the user, and thus attribute properties of both interaction and involved physical object as data of the user.

8.1.1. Augmented graphic tablet

The technical implementation of the augmented graphic tablet is similar to the augmented keyboard described in (Bovermann et al. 2010), replacing the keyboard by a rectangular plate with location tracking system. For the first prototype (see Fig. 8.4), we decided for a stylus-based graphic tablet which is basically a combination of both plate and tracking system. An advantage of the graphic tablet (compared to resistive sensitive surfaces) is, that the pen coordinates are already tracked while the pencil hovers contact-free over the surface, so that for instance filter parameters can be adjusted before any physical interaction happens. This is less critical for continuous interactions such as painting or scratching, but more so for tapping where the filter should have been parameterized *before* the pencil-paper interaction delivers the source signal for filtering. We use a Wacom Intuos 5 touch M which offers an active area of 244 mm × 140 mm or approximately the size of US Half Letter or A5 format, just enough for drawing simple sketches or writing small texts.

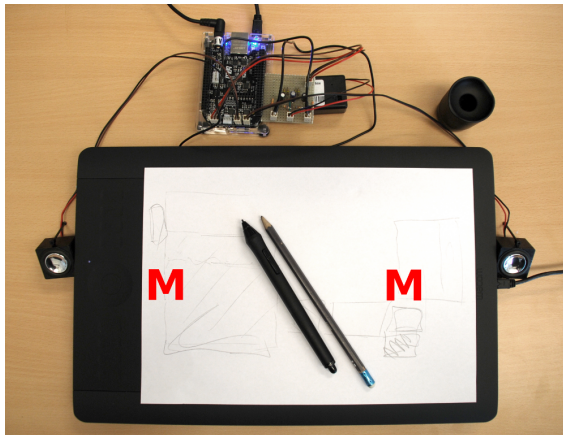


Figure 8.4.: The Mondrian graphic tablet. Microphone positions are marked by a red ‘M’.

As we know that spatial congruency between sensory modalities has an impact on plausibility (Yang and Yeh 2014), sound playback needs to be somehow spatialized in a meaningful way. Horizontally, a landscape oriented sheet of size A5 within reach for writing comfortably covers an angle of approximately 21° . This is well below the usual 60° for simple stereophonic panning strategies with two loudspeakers. Considering a perceptual localization blur of $\pm 3.6^\circ$ for white noise bursts in the horizontal plane (Blauert 1997, p. 41), a stereophonic setup is supposed to be sufficient for the given task. Due to the small size of the sonification interface, vertical spatialization is regarded as irrelevant, taking into account the stronger perceptual localization blur in vertical angle ($\pm 4^\circ$ for white noise, Blauert 1997, p. 44) as well as in distance perception ($\pm 25\%$ at 1.1 m for impulses, Blauert 1997, p. 47). Contact microphones and loudspeakers are therefore placed vertically centered on the left and right side of the graphic tablet to provide natural panning in the case that both signal paths (left and right) are processed individually.

The latency of augmented auditory feedback needs to lie below a certain threshold in order to be successfully and plausibly combined with visual or haptic information, as also emphasized by Müller-Tomfelde and Münch (2001) who referred to naturalism in this context. For trained users such as musicians, a latency below 10 ms is generally recommended (see Sec. 2.5.1).

As hardware platform, we therefore decided for the BeagleBone Black Rev. C with Bela audio cape, which is designed for sub-millisecond-latency au-

dio and sensor processing (McPherson and Zappi 2015) and targeting specifically for digital musical instruments (Zappi and McPherson 2014). Two low-cost piezo-electric contact microphones are installed on the left and right side underneath the graphic tablet’s active area. These are connected to the Bela cape through a FET buffer preamp (see Tillman 2005 for schematics), driven by a 9 V battery. Two miniature loudspeakers are connected directly to the on-board class-D amplifiers of the Bela, and are placed besides the tablet.

For sound synthesis, in contrast to the sonification of pen strokes described by Müller-Tomfelde and Münch (2001), we directly use the input signal from contact microphones as excitation signal for filter-based modal synthesis as described, e.g., in (Cook 2003, pp. 46–48). The physical model is intended to synthesize the normal modes (transversal waves) of a two-dimensional plate and to ‘apply’ these to the physical surface. The plate is seen as a linear and time-invariant system; its impulse response can be decomposed into exponentially decaying pure sine waves (Fletcher and Rossing 2010, p. 12):

$$x(t) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} A_{mn} e^{-\alpha_{mn}t} \sin(2\pi f_{d,mn}t + \phi_{mn}), \quad (8.1)$$

with mode indices m and n referring to the number of nodal lines in the two dimensions, respectively. Each mode is described by four factors: a frequency f_d , an attenuation rate α for the exponential decay, an amplitude A , and a starting phase angle ϕ , all dependent on mode indices m and n . In our simplified model, we assume amplitude and phase primarily influenced by the excitation signal and thus decided to ignore these in the first prototype.

As an example, we analyze an undamped, isotropic, and rectangular thin plate which is simply supported (hinged) along the edges. Its natural frequencies $f_{0,mn}$ are computed as follows (Fletcher and Rossing 2010, p. 80):

$$f_{0,mn} = 0.453 h \sqrt{\frac{E}{\rho(1-\nu^2)}} \cdot \left[\left(\frac{m+1}{l_x} \right)^2 + \left(\frac{n+1}{l_y} \right)^2 \right]. \quad (8.2)$$

In Eq. 8.2, the plate dimensions are described by width l_x , height l_y , and thickness h . The material of the plate is expressed by Young’s modulus E , Poisson ratio ν , and density ρ . The damped frequencies then become $f_d = \sqrt{f_0^2 - (\alpha/2\pi)^2}$.

8. Case studies of auditory augmentations

The damping of a plate depends on several factors including size, shape, material, and boundary condition, and cannot be easily predicted (see also Sec. 3.2.6). We therefore employ a simple but perceptually-tuned damping model that is parameterized by global damping α_G and frequency-relative damping α_R (Aramaki et al. 2011):

$$\alpha = e^{\alpha_G + \alpha_R \cdot 2\pi f_0} . \quad (8.3)$$

The resonant frequencies further depend on the boundary conditions of the plate edges. We describe the four boundary conditions through $B_{x0,i}$ and $B_{x1,i}$ for the left and right edge, and $B_{y0,i}$ and $B_{y1,i}$ for the top and bottom edge, respectively, where 0 means free, 1 means simply-supported/hinged, and 2 means clamped. Solutions for all combinations of these are derived by Warburton (1954).

In our simplified physical model, we neglect the resonant behavior of the graphic tablet and thus interpret the input signal from a contact microphone as a pure excitation signal induced by the stylus or pencil. This excitation signal is then filtered through our physical model to form the augmented output signal. Technically, this model is represented through a parallel filterbank of resonant band-pass filters which are tuned to the frequencies and attenuation rates of the simulated plate's normal modes. The implementation is done in SuperCollider (SC), based on the `DynKlank` object which adds any number of resonances to an input signal, taking vectors of undamped natural frequencies f_0 , amplitudes A (set constant to 1), and -60 dB decay times $T_{60} = \ln(1000)\tau$ (in seconds, with time-constant $\tau = 1/\alpha$) as arguments. The input signal is additionally cleaned from unwanted low-frequency noise through a high-pass filter set below the lowest mode frequency; the output signal is soft-limited.

Only the audio signal path is processed on the Bela running an SC server, while the mapping between data, physical model, and filter parameters is done on a separate computer running the SC language. Both computer and Bela communicate via LAN. The graphic tablet, connected via USB to the computer, is accessed as HID device.

In order to simplify the process of defining regions of different physical properties on the tablet already during development, a so-called Mondrian Generator¹ was used. This simple program generates compositions in red, blue, and yellow, "in the style

¹Mondrian Generator:

<https://github.com/JEFworks/mondrian-generator>

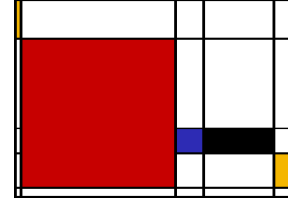


Figure 8.5.: Example image, generated with the Mondrian Generator, serving as test data.

of Piet Mondrian". The result is a raster graphics; an example is shown in Fig. 8.5.

Each of the three colors red, blue, and yellow represents a reference model M_i with specific physical properties, with i being the index of the model. Each model M_i is represented by its parameter vector:

$$\mathbf{p}_i = [B_{x0,i}, B_{x1,i}, B_{y0,i}, B_{y1,i}, E_i, \rho_i, \nu_i, \alpha_{G,i}, \alpha_{R,i}, l_{x,i}, l_{y,i}, h_i] \quad (8.4)$$

The three reference models are rendered in parallel with pre-set parameter vectors \mathbf{p}_1 , \mathbf{p}_2 , and \mathbf{p}_3 , respectively. Together they form the set of possible auditory augmentations. The color directly controls the input gains $g_{in,1}$, $g_{in,2}$, and $g_{in,3}$ of the three models, respectively (red: $g_{in,1} = 1$, $g_{in,2} = g_{in,3} = 0$; blue: $g_{in,2} = 1$, $g_{in,1} = g_{in,3} = 0$, yellow: $g_{in,3} = 1$, $g_{in,1} = g_{in,2} = 0$). White leads to a level-preserving mix of the three models. Black means that all input gains are set zero; black lines between the different regions are thus interpreted as grooves filled with sound-absorbing material.

For evaluation of the prototype system, we defined three reference models with plate dimensions l_x and l_y set to match the active area of the graphic tablet. All three models were set to a constant thickness h of 3 mm, with the boundary conditions for all edges set free, i.e., simulating plates that were freely hovering in the air. The material properties, as listed in Tab. 8.1, were set to match the material categories metal, ceramics, and glass, respectively.

Due to limited computing power on the BeagleBone, only few resonances can be processed. However, the number of vibrational modes in the relevant audible range, i.e., below 16 kHz, is anyway small for the synthesized plates, especially due to the small spatial dimensions. In particular, these were 13 modes for metal (starting at 1.5 kHz), 6 modes for ceramics (starting at 3.0 kHz), and 13 modes for glass (starting at 1.6 kHz). For each model, only the 6 lowest out of 100 calculated modes have been used.

Table 8.1.: Physical properties of the reference models.

	M_1	M_2	M_3
material category	metal	ceramics	glass
E (GPa)	180	360	70
ρ (kg m^{-3})	7740	3800	2600
ν	0.305	0.22	0.22
α_G	0.6	1.55	2.0
α_R ($\times 10^{-4}$)	2	1.75	1.5

For informal evaluation, an unknown random image was created by the Mondrian Generator and sonified through the Mondrian Table by using the pre-set reference models. The graphic tablet was covered by a blank sheet of paper. The given task was to trace back the sonified structures with the stylus, and to draw a visual representation of the auditorily perceived structure. For this task, the stylus was paired with a real pen in order to be able to paint contours on the paper while exploring borders of a region of equal material properties. Such a sketch can be seen in Fig. 8.4.



Source 8.1 shows a video demonstration that was recorded from first-person perspective with Soundman OKM binaural microphones (listening with headphones is recommended). It illustrates some tapping and scratching interactions with the pen at various locations of different data-driven (resp. location-dependent) augmentations.

The evaluation of the experimental platform provided us with valuable information on further development. In accordance with related studies, latency seems to be a major factor for plausible and successful auditory augmentation. During the evaluation, a round-trip latency of 25 ms was disturbing, while 14 ms was sufficient for the illusion of realistic auditory feedback when watching another person interacting (auditory-visual condition), but still felt unnatural for the interacting person (auditory-visual-haptic condition). With measured round-trip latency of about 1.5 ms, obtained with the Bela, the augmentation felt completely plausible and blended well with the direct sound of the original auditory feedback, despite spatially incongruent playback through loudspeakers. The synthesized materials felt realistic and successfully created the impression of a different materiality of the augmented physical surface. The sonified geometric shapes could be easily detected without effort while interacting in a completely natural way.

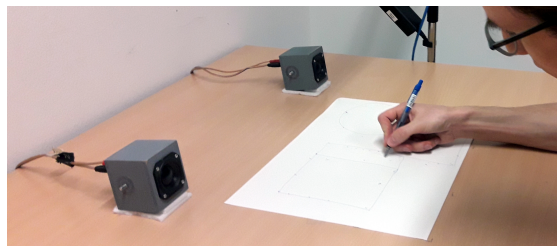


Figure 8.6.: The auditory coloring book.

8.1.2. Auditory coloring book

In order to get an impression on how unbiased and naive users might interact with such an augmented surface, we installed a modified version of the above-described system at an open house event of the university: the Auditory Coloring Book (see Fig. 8.6) is an interactive sound installation which auditorily displays regions of different color through augmented surface material of a real table. In contrast to the first prototype, this installation uses a real table equipped with two AKG C411 contact microphones underneath the surface. Two miniature loudspeakers are placed on top. Tracking of the pen is realized in Processing² through the Microsoft Kinect v2 sensor (see Sec. 4.7.1), covering a DIN A2-sized sheet of paper, fixated on the table.

During the event, volunteers were asked to trace back the regions of different materiality through natural interaction either with fingers or a ballpoint pen, both producing sufficient auditory feedback to drive the auditory augmentation. The pen additionally provides an intuitive way for switching between “exploration” and “drawing” mode through (de)activation of the tip.

Due to the lower resolution of the Kinect-based tracking compared to the graphic tablet, the fine-grained “Mondrian-style” structures were regarded as too difficult to detect in the envisaged context. Instead, users were asked to detect three different shapes, a rectangle, a triangle, and a circle, and to draw them on the paper sheet. The same materials as in the informal evaluation were mapped to the three shapes: rectangle to metal, triangle to ceramics, and circle to glass. The spatial arrangement of the shapes randomly differed between participants. The shapes could overlap, i.e., partially mask others. An example-image with the correct proportions was shown to the participants. Most of the time, the room was crowded with people, inducing a relatively

²Processing: <https://processing.org>

8. Case studies of auditory augmentations

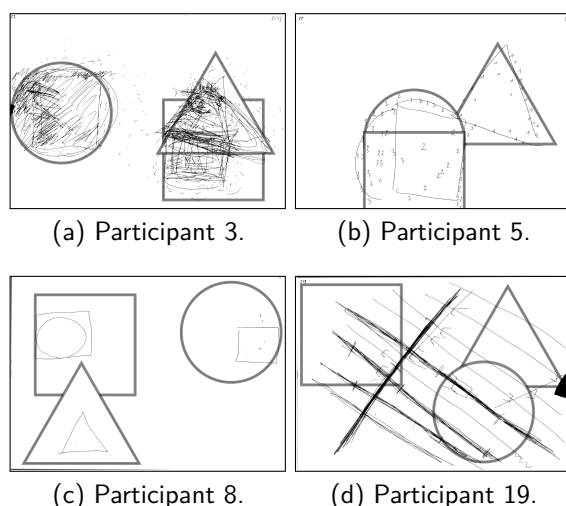


Figure 8.7.: Drawings from individual participants, overlain by the corresponding “correct answer”.

noisy and sometimes disturbing environment.

20 random visitors were documented through their individual drawings. Four of these drawings are shown in Fig. 8.7, with the auditorily displayed arrangement drawn as an overlay. An analysis of these drawings through visual judgment of the investigator reveals that 55 % of the participants correctly identified the approximate location of all three regions. 40 % additionally assigned all correct shapes, while 60 % roughly hit the correct size. 95 % confidence intervals, assuming binomial distribution, are [32, 77] %, [19, 64] %, and [36, 81] %, respectively.

Results indicate that the task was very hard to accomplish in the given context. However, the experiment clearly reveals four different strategies for task completion (compare Fig. 8.7): (1) continuous oscillatory drawing with enabled pen, somehow tracing back the border regions (top left), (2) tracing back the borders through systematic tapping, marking positions of different sound and connecting these (top right), (3) random tapping interaction to quickly find extreme values for efficient identification of the shapes (bottom left), and (4) continuous scanning of the surface and marking material changes (bottom right).

Yet interestingly, strategy (3) was only used by children and led to quick and reliable results for shape identification; however, always underestimating the size. The other strategies concentrated on exact border positions, and took much longer (up to about 15 min), but gained similar hit-rate for shape-identification. The relatively low overall

performance surely comes from low resolution and instability of the tracking system.

8.1.3. Discussion

With the described prototypes, we demonstrated the feasibility of two important aspects of auditory augmentations: (1) a high-fidelity control of discrete and continuous physical interactions, and (2) a manipulation of the auditory feedback between subtle and gross. These first prototypes have set the stage for more sophisticated auditory augmentations such as AltAR/table (Ch. 4).

8.2. “PilotKitchen”: display of electric load by virtual room acoustics

This section is based on the original publication which was created in teamwork by Groß-Vogt, Weger, Höldrich, Hermann, Bovermann, and Reichmann (2018). My main contributions include the experiment apparatus, load measurement, and data processing.



Tünnermann et al. (2013) demonstrated through their Knock'Knock system that variable room acoustics might be a promising channel for auditory augmentation. In this blended sonification, the reverberation of an empty office, perceived from the outside through knocking on its door, communicated the time that had been passed since the office was left. A visitor could thereby know from the length of the reverberation tail, if it would be worth to wait in front of the door, or rather try another day. While this system was intended to be used in absence of a person within the augmented room, we wondered how people would accept such a system while using the room that is being augmented.

Parthy et al. (2004) examined more systematically, how reverberation can be used for ambient data communication. They conducted a laboratory experiment in which participants had to identify the more reverberant stimulus in A/B comparisons. The stimuli consisted of music that had been processed by additional reverberation. They found that, in accordance with Weber's law, the reverberation time needs to be increased by approximately 60 % or decreased by approximately 30 % in order to be clearly perceptible. It must be noted that music is a rather difficult carrier signal for reverberation to be perceived, as it rarely exhibits pauses that allow a listener to estimate the length of the reverb tail. We assume that changes in reverberation get

more salient in the presence of signals that exhibit a better balance between transients and pauses. In addition, the use of not only decay time, but also amplitude of reverberation is assumed to increase salience and thus the amount of data that can be conveyed by the auditory augmentation.

The “PilotKitchen” was a pilot study for exploring the auditory augmentation of a kitchen via virtual room acoustics, with the goal to reveal problems that might occur in a practical application of this type of auditory augmentation. Therefore, we installed it at our university institute’s kitchen: a very specific surrounding that is neither private nor really public, with a large but limited group of users that use it on a daily basis, alone or for a chat between colleagues.

For sonification, we chose a type of data that affects everybody: electric power consumption. As being part of a public university, we are obliged to raise awareness as well as to develop solutions for the greatest challenge of our times: the transition to a sustainable life in order to avert the worst consequences of the human-made climate change. One key issue is a responsible use of energy. Energy, however, is still some kind of abstract entity that people generally have difficulties to conceptualize. In this sense, the augmented kitchen aims at making the use of energy perceptible by mapping it to the virtual room acoustics of the kitchen. The institute’s kitchen is especially interesting, as its users generally do not pay for the amount of energy they consumed, as would be the case in an in-home setting where the consumed energy may materialize in form of consumed money. While the system intends to raise energy awareness among colleagues, no actual change in behavior is expected within the scope of this experiment.

Around 15 colleagues who work within the building use the kitchen on a regular basis. In addition, several external lecturers, students, and visitors use it. From time to time, small meetings are held within the room. All users of the kitchen are generally familiar with virtual room acoustics.

According to the literature, successful feedback typically comprises appliance-specific data, presents the information in an appealing and intelligible way, and last but not least offers interaction (Abrahamse et al. 2005; Fischer 2007; Darby 2006). Schwartz et al. (2013) suggests that feedback should be specific for a certain task, and given continuously in real time. They argue that in the HCI community, ambient ecological feedback systems are mainly used for their motivational effects. Furthermore, the con-



(a) overview of the kitchen



(b) appliances with attached power meters

Figure 8.8.: The kitchen of the Institute of Electronic Music and Acoustics (IEM). Photos: Till Bovermann / IEM.

veyed information needs to be contextualized for the users, so that not only the pure physical entity is displayed, but rather its meaning in the given context.

8.2.1. Apparatus

The institute’s kitchen is a 16 m² room equipped with a table and 8 chairs and a kitchen unit including 5 appliances: dishwasher, coffee maker, water kettle, microwave oven, and fridge. A photograph is shown in Fig. 8.8a.

Electric load of individual appliances was measured by five Fibaro Wall Plug power meters which send data wireless over the z-wave protocol (see Fig. 8.8b). The sampling rate was 1 Hz, with measurements in the range between 0 and 3 kW. A laptop running Debian Linux was used for data processing and sonification. The data was received through an Aeotec Z-Stick USB dongle by using the

free and open-source Domoticz³ home automation system. A Pure Data patch forwarded the measurement data via OSC to SuperCollider where the data and audio processing was performed.

The input to the auditory augmentation system, comprising all sounds that were created in the kitchen, was captured by an AKG CK92 Blue-Line omnidirectional microphone hanging from the ceiling. The microphone was connected to the laptop via an M-Audio MobilePre audio interface. The outputs were connected to two Genelec 8020CPM loudspeakers that were mounted at the ceiling.

Additional sound absorbers were installed in the kitchen, in order to dampen it and bring down the natural reverberation time, so that also negative sonification (less reverberation than normal) was possible by the reverberation system. The aspects concerning the sound design are described in Sec. 8.2.4.

8.2.2. Electric load in the kitchen

In a first step, data has been recorded for several weeks, in order to get a first impression on the typical power cycles of each appliance, and on the usage patterns typical to the kitchen users.

For such appliances, Schwartz et al. (2013) discriminate between background services within the domestic environment, which consume energy without user interaction (e.g., fridge, and also dishwasher after it has started), and embodied services which are initiated by the user.

The individual appliances exhibit varying power cycles. The automatic coffee machine takes about one minute to grind the coffee, boil the water, and pour the coffee, and shows a very distinct load pattern. For the water kettle, the required heating time varies with the amount of water, with a binary load pattern (on/off). The fridge has a rather regular cycle of cooling. This load cycle is only weakly connected to user interventions such as opening the door. The dishwasher is embodied for the user who starts it, but more a background service for others. Once turned on (this happens between 2 or 3 times a week), it runs for about two hours. It consumes the most energy when it is silent: when heating up. The microwave is barely used, and has a binary load pattern similar to the kettle. All appliances also drain power in standby mode. One week of overall load is plotted in Fig. 8.9a.

³Domoticz home automation system:
<https://www.domoticz.com/>

The appliances' load cycles are overlain by the individual patterns of the kitchen users. Fig. 8.9a shows the presence of the institute's staff which reaches its weekly maximum at the institute's "jour fixe" on Tuesday afternoon, clearly visible by the peak in the measured electric power, mainly attributed to coffee consumption. On weekends, the kitchen is generally less frequently used; however, this usage is not fully reflected by the overall load, as large consumers such as the dishwasher generally dominate the load pattern.

8.2.3. Data processing

The auditory augmentation system for the institute kitchen was intended to convey two layers of information by using continuous auditory feedback. First, it should display the actual load, so that users could get immediate feedback on their actions such as switching on the water kettle. Second, it should relate the instantaneous load to the typical weekly load pattern. The data is therefore processed by an initial pre-processing stage and four iterative steps:

0. *Baseline initialization.* The system is initialized with data of one average week of overall load, which has been gathered during the three weeks' monitoring period (see Fig. 8.9b). This baseline contains $7 \text{ d} \times 24 \text{ h} \times 60 \text{ min} \times 60 \text{ s} = 604\,800$ samples that represent each second of the week.
1. *Smoothing.* Most appliances exhibit discontinuous load patterns due the binary nature of their components (e.g., heating, cooling, coffee grinding, etc. are just switched on and off). The raw measurement data is therefore smoothed by a leaky integrator:

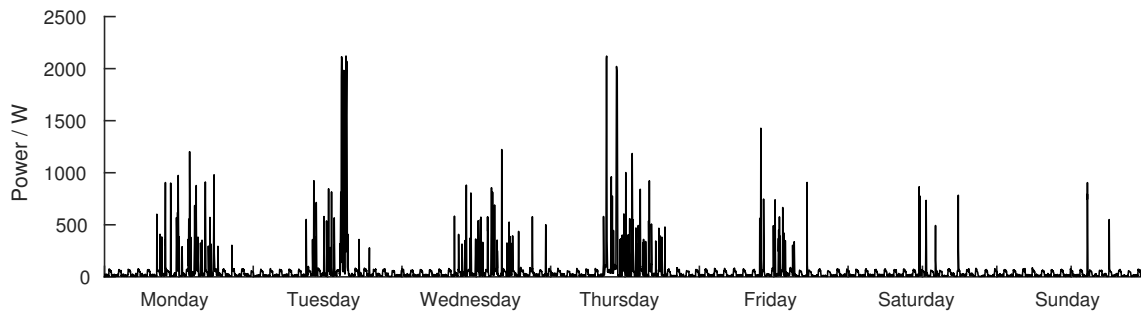
$$y[n] = (1 - a)x[n] + ay[n - 1] \quad (8.5)$$

with discrete input signal $x[n]$ at sample n and output signal $y[n]$. The smoothing constant a describes the amount of smoothing, and relates to the time constant τ and sampling interval T (here 1 s) via

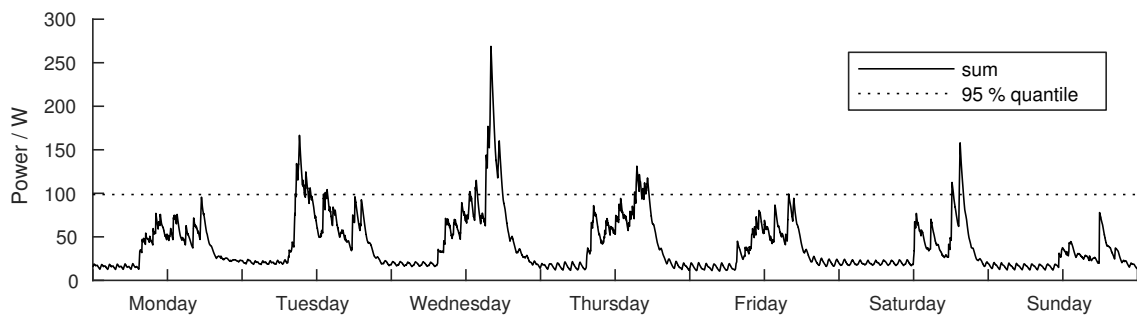
$$a = e^{-T/\tau} \quad (8.6)$$

A suitable value for τ has been empirically adjusted in order to attain a good balance between sufficient smoothing and adequate temporal resolution for real-time feedback. In experiment 1 it was set to 60 min. In experiment 2, τ was lowered to 15 min, which means that after the period of 5τ , past data has decayed to less than 1 % of its initial value.

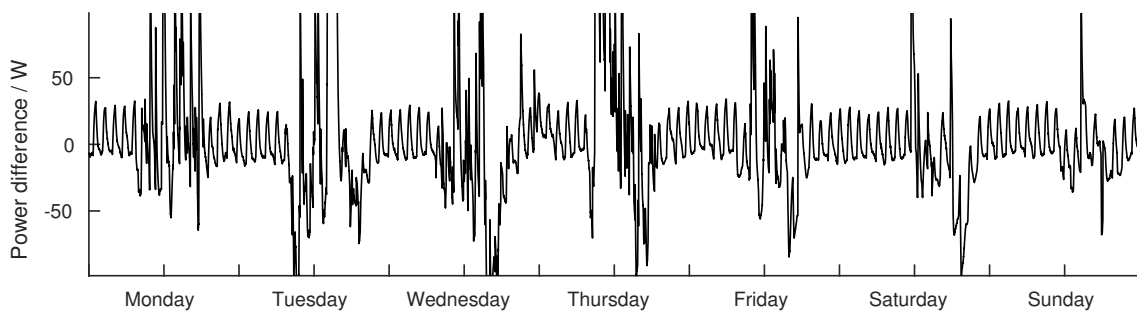
8.2. "PilotKitchen": display of electric load by virtual room acoustics



(a) Raw load smoothed with $\tau = 1$ min.



(b) Baseline load, averaged over three weeks in autumn 2017, smoothed with $\tau = 1$ h.



(c) Difference between the raw load, smoothed with $\tau = 15$ min, and the baseline.

Figure 8.9.: One week of electric power consumption during the second evaluation study, 12–18 March 2018.

8. Case studies of auditory augmentations

2. *Baseline update.* For each new time-step, the baseline is updated according to the values of the past weeks, by using another leaky integrator with $a = 1/3$. This means that the actual smoothed power is iteratively averaged with the value of the past weeks with a weight of one third.
3. *Power difference.* For each time-step (i.e., second), the smoothed power is subtracted from the weekly baseline, resulting in the pattern shown in Fig. 8.9c for one week. The function is positive if the smoothed (i.e., typical) consumption is above the baseline, and negative if it is below.
4. *Clipping.* The power difference is clipped along the 95-percentile of the baseline data, so that statistical outliers are eliminated. The resulting value serves as the driving signal for the sonification.

The steps 1 to 3 are updated with every new measurement value, every second. Step 4 is updated once a week, as no large variation of the 95-percentile of the baseline is expected.

8.2.4. Sonification design

The sonification design basically maps higher load to a more reverberant room. It is implemented in SuperCollider, with a reverb based on the JPVerb plugin⁴. The relevant parameters of the reverb are: *t60* (−60 dB reverberation time), *damp* (damping of high frequencies between 0 and 1), *size* (the simulated room size), and *high* (multiplier of *t60* above the cutoff frequency *highcut*). Based on these parameters, we set three distinct *presets* which define the parameter space:

0. *zero*: no augmentation, i.e., no added reverberation (parameter settings as in preset 1, but with zero gain.)
1. *kitchen*: additional reverberation which is so natural or plausible that only changes from or to this preset are perceived.
Settings: [*t60*: 0.8, *damp*: 0.0, *size*: 1.1, *high*: 0.3, *highcut*: 1500]
2. *church*: the reverberation clearly exceeds plausible kitchen room acoustics, both in reverberation time and level.
Settings: [*t60*: 3.0, *damp*: 0.6, *size*: 2.0, *high*: 0.5, *highcut*: 1000]

⁴SC3-Plugins:
<http://supercollider.github.io/sc3-plugins/>

The driving signal for the sonification, as described in the previous section and plotted in Fig. 8.9c, is mapped to the sonification parameters so that a value of −1 leads to preset 0 (zero), a value of 0 leads to preset 1 (kitchen), and +1 leads to preset 2 (church). In between, all parameters are interpolated linearly. The room acoustics is thus constantly changed by the overall load, except for the extreme values that are clipped. For negative values (relatively low load), the sonification produces a plausible amount of reverberation, and thus a rewarding sound. For positive values (relatively high load), the reverberation gets implausible, up to a level that might even be disturbing for the users in some situations where speech intelligibility is important. The presets were tuned so that, from the designer's perspective, small deviations from the baseline (kitchen preset) can be easily perceived.

The sonification is designed to display two kinds of information at the same time. First, it displays changes in overall load through relative changes of the reverberation. It thus gives immediate feedback on the instantaneous change in load. Second, the sonification displays the difference of the instantaneous load to the baseline, i.e., to the average load at this time and day of the week, through the absolute amount of reverberation. It thus relates the abstract absolute load to the average value.

A demo video of the running system from a first-person perspective with binaural sound is available in Source 8.2.



8.2.5. Evaluation

The sonification was evaluated in two pilot experiments, aiming at the perceptibility and also contextualization of the conveyed information, respectively. The participants of both experiments were the 15 colleagues (3 female, 12 male) who work in the same building and use the kitchen on a regular basis. While these participants are surely not representative for the general population, they can be regarded as experienced listeners with a background in sound engineering, music, or sound design.

8.2.5.1. Perceptibility

Within the first round of evaluation, the system ran for ten days during the end of the winter semester — a busy period in which the kitchen is frequently used. The members of the staff were informed beforehand that an experiment would be conducted in the kitchen, involving sonification of electric load.

However, no further information was given to the users.

After ten days, a questionnaire was distributed among participants, in order to collect impressions on the sonification system. It contained standardized and open questions on the participants' use of the kitchen and on their assumptions and opinions on the sonification system. Eleven participants returned the filled questionnaire.

Concerning the usage of the kitchen, 7 participants indicated to use it 3 to 5 times a week, 3 participants indicated to use it 1 to 2 times a week, and only one participant stated to use it less often.

Concerning perceptibility, 9 of the 11 valid participants indicated to have noticed something in the kitchen during the test period. In addition, these participants all correctly identified resonant sound phenomena, howling, or a changed soundscape. While this might show that the majority of participants perceived the augmented auditory feedback, it also shows that it might not have been perceptible all the time. Only three participants indicated that they had noticed something on all occasions, while in addition, two of those were not sure about it. These results suggests that the sonification reached its goal of at least partially staying below the threshold of perception, e.g., in case of an instantaneous load below the baseline.

Some individual answers gave a hint on the calmness of the sonification. Two participants indicated that the feedback was subtle, and that the kitchen sounded more lively. On the other hand, three participants complained about the sound, and two even turned off the system during a joint meeting, by using the emergency switch that was originally only included for the case that something goes wrong.

Concerning interactivity, three participants reported that they felt able to interact with the sound over speech or by making other noises, even though they were not explicitly asked about it. One participant, however, found that his behavior had no influence on the sound.

Concerning the underlying information on electric load, none of the participants had worked out a hypothesis on the mapping between the sonification and the measured load.

These results of experiment 1 showed the need for two adjustments of the sonification system. First, the time constant for the smoothing of instantaneous load was lowered, in order to give feedback more promptly. Second, the preset of the upper threshold (church) was tuned in order to give a more pleasing sound quality, even at high load with

respect to the baseline.

8.2.5.2. Contextualization

The sonification was evaluated in the second experiment, incorporating the adjustments based on the results of experiment 1. The previous experiment also showed that the participants' individual statements could be arranged in a continuous space defined by positive or negative valence, as well as by the level of perception or control (from imperceptible, via passively perceptible, to interactive). This observation led to the idea of evaluating these subjective dimensions more systematically. As proposed by Soares et al. (2013), we evaluated the dimensions of valence, arousal, and control/dominance, based on self-assessment manikins (SAMs, M. M. Bradley and Lang 1994) which involve a graphical representation of the corresponding scale values between 1 and 5. The SAM technique is widely accepted for quick and non-verbal evaluation, especially in the context of sound.

Prior to the experiment, the system was explained to the participants, in terms of both data processing and the resulting auditory feedback. The experiment itself then involved a diary study, in which participants were asked to write a diary entry each time they used the kitchen, within the two weeks evaluation period. For that purpose, the prepared sheets of paper including a simple questionnaire were laid out, to be filled and submitted to a mailbox.

The diary questionnaire included some basic meta-data:

- participant's name or anonymized but consistent nickname
- date and time
- duration and reason of the stay in the kitchen

as well as questions concerning the soundscape, in case it attracted any attention:

- description of the situation
- 5-point self-assessment manikins (SAMs) for the dimensions valence, arousal, and control/dominance; measuring the affective reaction of the user towards the system
- any other reactions

In total, 14 participants actively took part in the survey. Six of them submitted diary entries on a regular basis, with 13.6 total entries per person on average. The other 8 participants handed in only three or less entries. Not all answers could be taken

8. Case studies of auditory augmentations

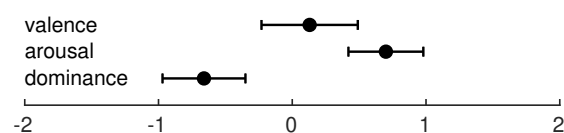


Figure 8.10.: Average ratings and 95 % confidence intervals across participants, for the dimensions of valence, arousal, and dominance, respectively.

into account, as the users occasionally deactivated the system (and sometimes forgot to turn it on again).

Overall, about half of the diary entries documented that participants perceived something. Surprisingly, however, these data showed no significant correlation with the driving value of the sonification. The hypothesis that the sonification would be easily detected above the baseline threshold could therefore not be supported by the data. This might be explained by the small sample size, as well as by the large interpersonal differences among participants. Some participants seemed to be very sensitive towards the sonification, e.g., describing the reverberation as “strong” or “reverberation of a church” even for low values just above the baseline. Other participants did not notice anything until much higher values.

The above findings could be supported by the SAM ratings which also differed strongly between participants. While the affective reactions were generally positive, some participants varied their ratings from entry to entry. Figure 8.10 shows the average ratings and 95 % confidence intervals across participants, for the dimensions of valence, arousal, and dominance, respectively. Generally, positive and negative attitudes towards the system were balanced: the average of valence was not significantly different to 0 (mean and 95 % confidence interval: 0.13 ± 0.36), at a significance level of 5 %. Furthermore, the system had a slightly arousing effect on participants (0.70 ± 0.28), but participants were generally passive about it (mean dominance: -0.66 ± 0.31).

8.2.6. Conclusions

The augmented kitchen was a first prototype for the ambient display of overall electric load by auditory augmentation in an ecologically valid setting. The load was mapped to artificial reverberation of the room, giving (a) immediate feedback on instantaneous load, and (b) relating that load to the average

load on the same time and day at previous weeks.

This prototype was designed to be simple and neutral, completely data-driven, and not judgmental with respect to an absolute level of consumption. At the same time, it is self-adaptive, while constantly updating the weekly baseline.

The two rounds of evaluations revealed several problems. For many participants, the system was not salient enough to convey the underlying information. This might be attributed to the stationary noise that is emitted by some of the kitchen appliances such as water kettle, dishwasher, or coffee machine. These sounds may have masked the users’ own interaction sounds which usually provide the best input with respect to salience of reverberation: short transient sounds that come, for instance, from footsteps or from placing objects on the table. A solution to this problem might be to adapt the level of reverberation dynamically to the level of the (machine-made) soundscape.

For the prototype and evaluation, we did not account for the load of the sonification system itself. The economization of electric energy that is achieved by the raised awareness must be higher than the energy consumption of the sonification system itself, in order to achieve a net energy saving. In a practical application, the system must therefore be designed for low energy consumption, e.g., by replacing the laptop with a low-power microcomputer.

In general, the participants of the evaluation did not feel to have much influence on the sonification. This might come from the fact that some user actions such as coffee making produced only little load in comparison to some high-power background services such as the dishwasher. We assume that this problem could be overcome if background services (e.g., dishwasher or fridge) were treated differently than embodied services such as preparing tea or coffee. One solution might be to apply stronger smoothing on appliances such as the fridge with its regular and binary load cycle that is (almost) uncoupled to user actions. A more radical solution would be to not sonify background services at all, as users anyway have no choice to change their behavior on these machines, except buying more efficient ones.

The real-time nature of the sonification, in relation to the past load of this day and time of the week, implied that the load in times of absence is never perceived by the users. This period, however, is crucial for saving energy. At least for the institute’s kitchen, the largest factor for energy economization

is to turn of appliances during the night. In future sonification designs, this factor must be included.

The overall setting was a very specific one that cannot be compared to households. Conclusions on any hypothetical positive effect of the developed sonification on domestic energy economization could therefore only be made if it were tested in a real in-home setting.

Concerning the use of artificial reverberation for auditory augmentation, the augmented kitchen also brought up some general questions regarding perceptibility. How much information can actually be conveyed by reverberation in an ambient sonification setting? And how many different levels of reverberation can we absolutely identify (and thus map to the underlying data)? In order to shed light on these questions, we designed a follow-up experiment that is described in Sec. 8.3.

8.3. “RadioReverb”: room reverberation as ambient communication channel



This section is based on the original publication which was created in teamwork by Groß-Vogt, Weger, Frank, and Höldrich (2021). My main contributions include the experiment app and the statistical analysis.

RadioReverb was originally designed as a laboratory experiment under controlled conditions within the university facilities. Due to the Covid-19 pandemic, however, the experiment was re-designed towards a smartphone app, so that participants were able to perform it at home by using headphones. The main goal of the experiment was to examine, how much information can be conveyed at the periphery of attention, by auditory augmentation through virtual room acoustics. This implies that users perform a main task that, at the same time, should not be disturbed by the sonification. This experiment is also described in detail by Groß-Vogt et al. (2021).

The test design follows a magnitude estimation approach. Participants are asked to estimate the amount of reverberation on a given arbitrary scale, represented by a slider with a fixed minimum and maximum value. While in the described experiment the participants obtained prior training on the relationship between the slider value and the resulting amount of reverberation, no actual number such as reverberation time T_{60} are mentioned, as the general user is anyway not familiar with such acoustical measures.

8.3.1. Participants and their main task

We are not interested in the mere physiological limits of perception of reverberation, but rather in the human capabilities to monitor reverberation as part of a secondary task. Therefore, a reasonable main task is needed that attracts attention. In addition, the auditory augmentation through virtual room acoustics works only if there is some kind of excitation signal that makes the reverberation audible. According to Hazlewood et al. (2011), the displayed data in an ecologically valid experiment should be of interest to the participants. As our data is actually just random noise, we intended to at least provide an interesting main task: listening to the radio. This is an exceptionally simple main task that provides a controlled excitation for the reverberation. We chose two 15 min topical episodes of the radio feature “Ö1-Radiokolleg”, produced by the Austrian broadcasting corporation (ORF), on the topic of food plants (fig and millet, respectively), with the expectation that they would be (a) unknown and (b) of interest to the participants. Concerning the audio content, they comprised a good balance between raw speech in a dry studio and interviews or sound bites recorded outside the studio.

A total of 21 participants were recruited for the main experiment; however, only 17 of these (13 male, 4 female) were able to perform it until the end and to submit their answers. These are regarded as the experiment group (EG), with distinction to an additional control group (CG) comprising 12 participants (6 male, 6 female) who performed the main task only, i.e., listened to the radio features without auditory augmentation. The EG participants included 2 professional musicians with all others being staff or students of the institute. They all can be regarded as expert listeners with presumably normal hearing. The CG participants were recruited from among the acquaintances of the authors, and generally had no specific background in audio or music.

8.3.2. Stimuli: artificial room reverberation

8.3.2.1. A plausible range of reverberation

For the experiment, we assumed a virtual room of $4\text{ m} \times 5\text{ m} \times 3\text{ m}$ dimensions. As we target a practical in-home application, the artificial reverberation must stay within plausible limits of this setting. In the average apartment, the lowest amount of

reverberation is usually found in the living room or bedroom, containing lots of damping materials such as curtains, cushions, and upholstered furniture. While there are usually still room reflections present, the reverberation time may get close to zero. In contrast, a bathroom with tiled walls and floor usually exhibits the most reverberation of the whole apartment. Based on the corresponding physical constraints, we simulated virtual room acoustics within this plausible range, leading to a diffuse reverberation with decay time T_{60} between 0 s and 1.2 s. Details on the simulation follow in Sec. 8.3.2.2. In informal listening tests, these were considered as plausible room acoustics that may occur in the average apartment. This range was therefore selected as the parameter range for the sonification.

8.3.2.2. Binaural rendering of virtual room acoustics

The virtual room acoustics was simulated by using the IEM Plug-in suite⁵. Within and between the IEM plug-ins, the directivity of sound is encoded by using Ambisonics technology (see, e.g., Zotter and Frank 2019).

For simulating the directivity pattern of a loudspeaker (i.e., a consumer radio set), the monophonic input signal was fed into the DirectivityShaper plug-in which created a 3rd-order (16 channels) spherical harmonics representation of the typical frequency-dependent directivity. The SceneRotator plug-in was used to rotate the virtual loudspeaker so that it faced the listener from the front.

Early reflections were rendered by the RoomEncoder plug-in, which was set to recreate 236 distinct reflections of the scenery, based on the room dimensions (5 m × 4 m × 3 m) and the positions of the virtual loudspeaker at (1.7, 0.3, -0.5) m and listener at (-0.3, 0.3, -0.5) m relative to the center of the room. The resulting 2 m distance between listener and radio represents an ecologically valid situation. The reflection coefficients of walls, ceiling, and floor were set to plausible values of a living room with a carpet on the floor. Diffuse reverberation was simulated through a 64 × 64 feedback delay network (FDN) by using the FDNReverb plug-in.

The resulting Ambisonics stream was then rendered for binaural headphone presentation by using the BinauralDecoder plug-in. The plug-in incorporates state-of-the-art decoding technology (Schörkhuber et al. 2018; Zaunschirm et al. 2018);

the output signal can be regarded as perceptually similar to the hypothetical true signal, at least for the envisaged setting of listening to the radio.

The parameter settings that were used in the experiment are available for download in Source 8.3. </>

When interpolating between the two extreme settings, we had the impression that we could clearly discriminate between about 7 discrete steps. The perception of the corresponding sound parameters, however, needs a more detailed examination.

8.3.2.3. Discriminable levels of reverberation

Impulse responses of the simulated extreme settings (living room and bathroom) and in-between values allowed the computation of acoustical descriptors such as reverberation time or direct-to-reverberant ratio. The perception of such specific aspects of reverberation have been extensively studied in the literature, and therefore allow the estimation of the number of discriminable levels based on the number of JNDs that fit within the specific parameter range.

Niaounakis and Davies (2002) measured an average JND in reverberation time of 0.042 s, for short absolute reverberation times below 0.6 s. Above that threshold, we can assume the generally accepted JND of 5 % that was measured by Seraphim (1958). The resulting average absolute JND between 0.6 s and 1.2 s is actually on par with the one obtained by Niaounakis and Davies. The measured reverberation times between 0.07 s (reflections only) and 1.20 s (reflections and diffuse reverberation) therefore fit about 27 discriminable levels. As decay time is not the only sound parameter that changes between our two extreme settings, its perception in isolation might not reflect the actual perception of the virtual room acoustics.

The direct-to-reverberant energy ratio (DRR) yields only 4 discriminable levels, based on the measured JNDs by Larsen et al. (2008). In case of center time (Cox et al. 1993) as well as for clarity C_{50} (J. S. Bradley et al. 1999), a total of 11 JNDs fit in the observed range.

There is another aspect that needs to be taken into account besides mere perceptual differences: the number of distinct levels a human listener can actually memorize. Miller (1956) gathered data on the channel capacity for absolute judgments of parameters in various sensory modalities. His results consistently suggest a channel capacity that is equal to “the magical number seven, plus or minus two”. As this number is actually in accordance with our personal impression of discriminable steps from

⁵IEM Plug-in suite: <https://plugins.iem.at>

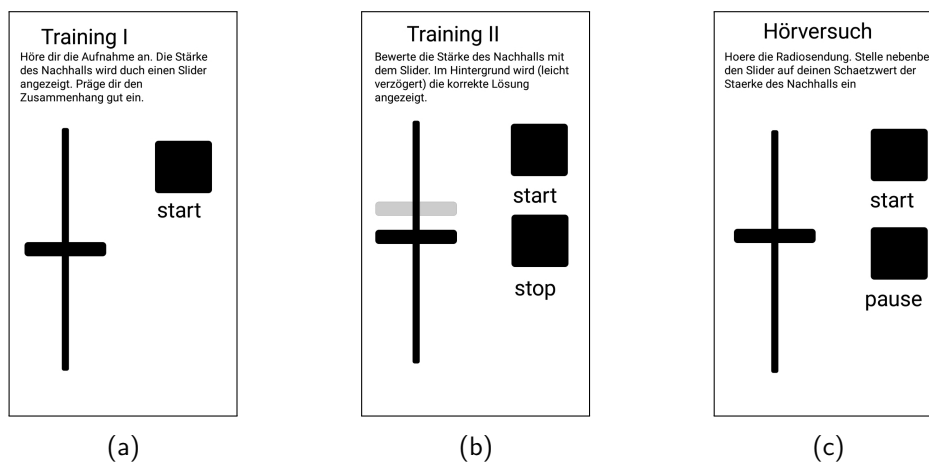


Figure 8.11.: The GUI of the smartphone app during the three phases of the experiment. (a) Training phase I: “Listen to the recording. The slider displays the amount of reverberation. Try to memorize its relationship to the sound.” (b) Training phase II: “Listen to the recording and assess the amount of reverberation on the slider. The correct value is displayed in the background after a short delay.” (c) Test: “Listen to the radio broadcast while constantly estimating the amount of reverberation by adjusting the slider.”

the informal listening test, we chose to use these 7 distinct levels of reverberation for the auditory augmentation within the experiment.

8.3.3. Apparatus: a mobile experiment app

We used the free MobMuPlat⁶ app by Iglesia (2016) in order to allow participants to conduct the experiment themselves at home on their smartphones or tablets. MobMuPlat is a “mobile music platform” for Android and iOS devices that provides a GUI for the audio engine in Pure Data (Pd). The experiment procedure itself was coded in Pd and controlled via the GUI which is shown in Fig. 8.11. The app played back the pre-rendered sound files which were individualized for each participant.

The participants returned their recorded response data via email in the form of a text file containing the slider values sampled at a rate of 100 Hz. Three of the 17 participants accidentally performed the experiment with the app set to the wrong audio sampling rate: participant 16 listened to the 44.1 kHz version at 48 kHz, while for participants 42 and 44 it was the other way around. The resulting change in playback speed (−8 % or +9 %) was considered to have no significant effect on the results, as JNDs for diffusivity and reverberation time are generally

⁶MobMuPlat:
<https://github.com/monkeyswarm/MobMuPlat>

defined relatively, and not in terms of absolute values. All collected data were resampled to the same length of 163 662 samples (about 27 min 17 s).

8.3.4. Procedure

8.3.4.1. Training and main experiment

The participants performed the experiment at their homes on their own smartphone or tablet devices with attached headphones (not earbuds). They were required to install the MobMuPlat app and load their personalized preset that was provided in form of a download link. Participants were asked to take a seat at a comfortable place in a normal room with little distraction. They were guided through the 3 stages of the experiment (training I, training II, and test, as shown in Fig. 8.11) as well as to a post-hoc survey, by following the instructions in the app.

In training I, participants were (passively) familiarized with the sound of the reverberation and the mapping between the amount of reverberation and the graphical slider (see Fig. 8.11a). A 46 s excerpt from a radio feature was played, while the amount of reverberation iterated through the whole parameter range in 12 discrete levels, starting from minimum, up to maximum, and back. At the same time, the slider in the GUI displayed the current true value. Participants were able to repeat the process at wish.

8. Case studies of auditory augmentations

In training II, a 3.5 s speech sample was played back in a loop, where every repetition had a randomly different level of reverberation. Participants were asked to (actively) estimate the level of reverberation by setting the slider. During the 1.5 s pause between repetitions, the true value was revealed by an additional slider in the background (see Fig. 8.11b), so that the participants were able to improve their performance. Training II used the same 12 levels of reverberation as training I. They were presented in consecutive rows which contained each level once in random order. The participants had to perform this procedure for at least two such rows (24 stimuli) in order to be able to proceed to the test.

In the actual test, the participants listened to the two radio features (played one after another without pause). At the same time, they were asked to respond to any perceived change in reverberation by setting the slider to the appropriate value as soon as possible (see Fig. 8.11c). This time, no feedback was given. The slider always remained at the last value that was set by the participant.

During the whole testing period, all possible transitions between the 7 levels of reverberation (i.e., at least 42 transitions) were presented in random order, different for each participant. Each level was kept constant for at least 7 s.

8.3.4.2. Post-hoc survey

After completion of the experiment, participants were asked to fill an online survey, comprising a multiple-choice test on the content of the radio features, as well as general questions concerning the participants' listening environment and their subjective experience with the sound:

- 14 multiple-choice, multiple-response statements on each of the two radio features (fig and millet). While all statements were factually correct, the participants had to select those that had actually been part of the radio feature.
- an open question where participants described the room and setting in which the experiment was performed, as well as possible technical issues.
- "How well do you think you were able to estimate the reverberation level?"
- "How much were you challenged by simultaneously listening to the radio and estimating the reverberation level?"

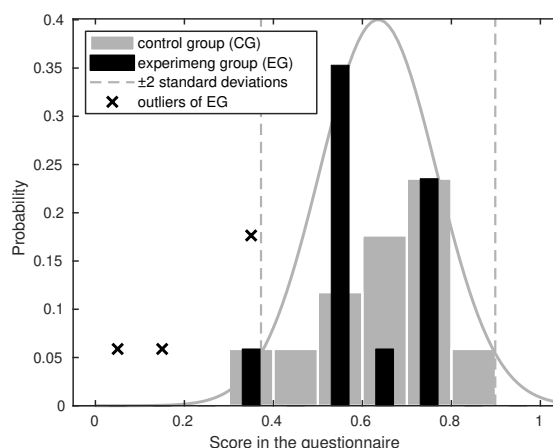


Figure 8.12.: Histogram of scores from the multiple-choice questionnaire, for EG and CG. Five outliers of EG had scores of less than 2 standard deviations below the average of CG, and were thus classified as sound-focused (SF) participants.

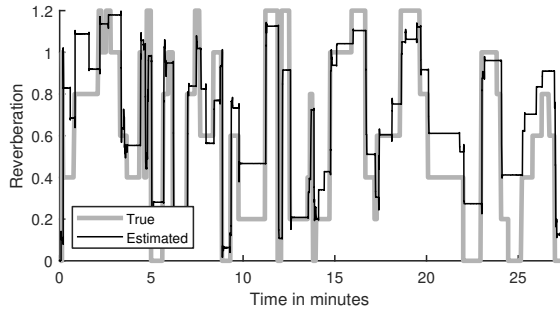
- "How realistically did the virtual radio and its reverberation blend into your listening environment?"

The last three questions had to be answered on an ordinal scale between 1 and 7. The participants of the control group (CG) answered only the first part concerning the content of the radio features.

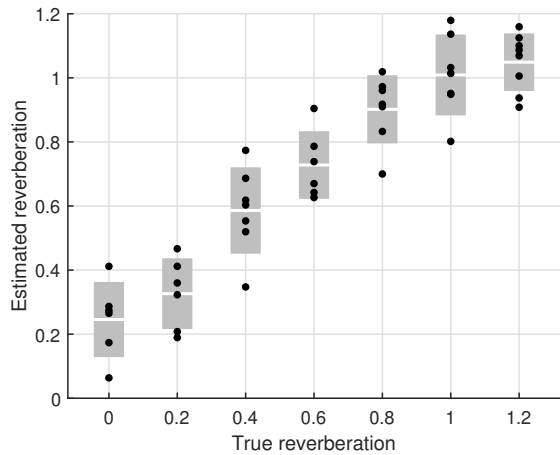
8.3.5. Results

First, the data from the multiple-choice questionnaire were analyzed. Fig. 8.12 shows the distribution of the EG data in comparison with that of the CG data. When we excluded the data from the five worst-performing participants, the results of the questionnaire from the CG were not significantly different from those of the remaining participants of the EG ($p=0.33$). When we compared the results of the CG with those of the EG, we found that some of the EG participants were much less focused on listening to the radio than the other EG participants. We call this group the sound-focused listeners (five participants with IDs 4, 6, 11, 23, and 42), whereas the other EG participants behaved as content-focused listeners in a more "peripheral" way. EG participants were therefore subdivided into sound-focused (SF) and content-focused (CF) listeners.

The data obtained from the MobMuPlat app were time series of estimated reverberation levels



(a) Time series of true and estimated reverberation.



(b) Estimated vs. true reverberation, with mean values and standard deviations per level.

Figure 8.13.: Experiment data of participant 16.

(response) in relation to the true reverberation levels (reference) for each participant (see example in Fig. 8.13a). In order to simplify comparisons, both are given in quantities of reverberation time T_{60} . We estimated the delay of the response with respect to the reference via the maximum of their cross-correlation. This average delay was between 0.54 s and 2.22 s for all participants (mean: 1.48 s, standard deviation: 0.55 s). A moving-average instantaneous delay over time was computed in the same way by using a sliding rectangular window of four minutes length. This instantaneous delay ranged from 0 to 3.88 s for all participants, with an average delay over time of 0.99 s.

For further analysis, the time series were divided into segments of constant true reverberation (the plateaus are shown in Fig. 8.13a in the gray reference curve). The first segment was excluded. In addition, the first 2.22 s (i.e., the maximum average delay), as well as the last second of each segment were removed; for the remainder, the average re-

sponse value (i.e., the estimated reverberation) was calculated. After collecting average responses, together with the corresponding reference levels, we obtained between 6 and 9 estimates per participant and reverberation level, as shown in Fig. 8.13b.

For each participant and level, the distribution of estimated reverberation was tested for normality by the Lilliefors test. In only 4% of the cases (and for a maximum of one level per participant), the null-hypothesis of normal data was rejected. Therefore, we generally assumed a normal distribution for further statistical analysis. Within each participant, we performed pairwise, one-tailed, Welch’s t-tests on all possible pairs of reference levels, with a 5% threshold for significance. For all participants, a difference of three reverberation levels (e.g., 5th vs. 2nd level) always resulted in estimates that differed significantly from each other. When pooled over participants, all pairwise comparisons were significant; i.e., the estimated reverberation of each level was always significantly higher than that of all levels below, and significantly lower than the levels above, respectively.

When analyzing estimated vs. true values, several different approaches can be used to define the number of levels that participants were able to identify correctly. As can be seen in the example of Fig. 8.13b, it was quite obvious that the participants did not reach our seven reverberation levels. Due to the small number of data points per level and the fact that each participant has his/her own non-linear mapping, we decided to use the statistical measure of effect size to analyze the data. Note that this choice follows the assumption that our levels are equally spread and that the perceptual distance between adjacent levels is similar over the whole range of reverberation.

Based on the effect size, the number of discriminable levels are computed in the same way as already described in Sec. 5.2.5.6. For some interesting values of threshold effect size d_t or the corresponding probability of superiority P_s , the results are given in Tab. 8.2 and Fig. 8.14; averaged over all participants, as well as averaged, separately, over the groups of SF and CF listeners. In Fig. 8.14a, N is plotted against a continuous P_s . When choosing a d_t of 2.77 (corresponding to $P_s = 0.975$), content-focused (CF) listeners achieved an average of 3.0 discriminable levels (standard deviation SD 0.4), while sound-focused (SF) listeners achieved an average of 3.5 levels (SD 0.4). The number of perceived levels was significantly higher for SF listeners than for CF listeners ($t(9.44) = 2.161$, $p = 0.029$).

8. Case studies of auditory augmentations

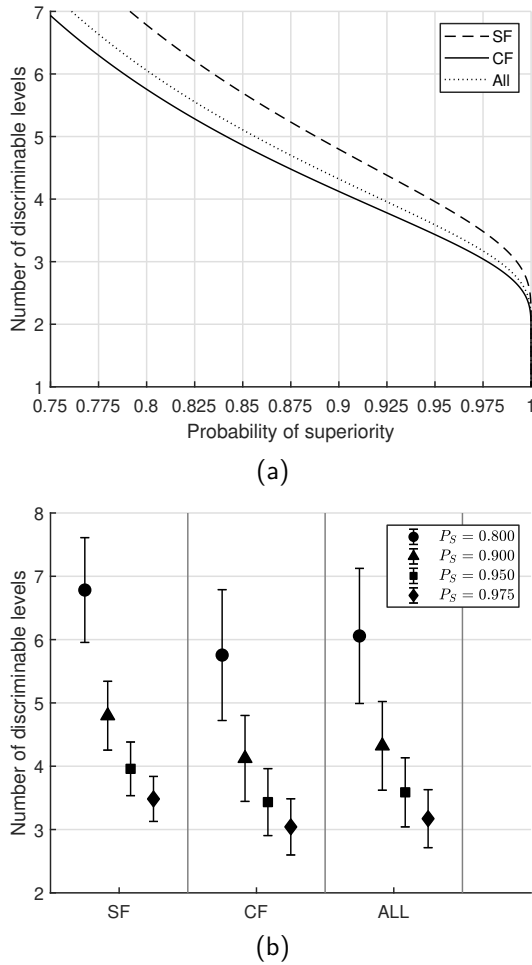


Figure 8.14.: Number of discriminable levels for different thresholds d_t (Tab. 8.2) and their corresponding probability of superiority (P_s). (a) depicts the number of levels as a function of P_s . (b) shows the same information for 4 selected values of P_s , together with the standard deviation across participants.

For each reverberation level, the individual estimates can be grouped into upward jumps (the prior reference level was lower than the current one) and downward jumps (the prior reference level was higher than the current one). In this way, the estimated reverberation can be plotted against the true reverberation for upward and downward movements separately. Due to the small number of data points, a within-subject analysis would not be meaningful. Fig. 8.15 shows the resulting curve, pooled over all participants. It shows two effects that can be expected when dealing with magnitude estimation (Petzschner et al. 2015). First, the regression effect

Table 8.2.: Probability of superiority P_s corresponding to the thresholds d_t and results for the number of levels discriminable by all, CF, and SF listeners.

P_s	d_t	N_{all}	N_{CF}	N_{SF}
0.8	1.19	6.1	5.8	6.8
0.9	1.81	4.3	4.1	4.8
0.95	2.33	3.6	3.4	4.0
0.975	2.77	3.2	3.0	3.5

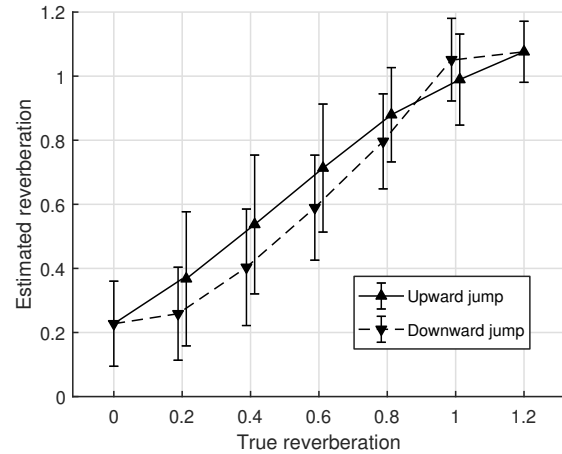


Figure 8.15.: Estimated vs. real reverberation, pooled over all participants, split into upward and downward level changes.

is the tendency that estimates are systematically biased towards the center of the distribution, i.e., the objective range of 0 to 1.2 is mapped to approximately 0.2 to 1.1. Second, the judgments depend on the recent history of stimuli, known as sequential effect. Estimates after a large previous stimulus tend to be larger, while estimates after a small previous stimulus tend to be smaller, leading to a perceptual hysteresis curve that is clearly shown in Fig. 8.15. Note that for levels 1 and 7 only one direction exists. For levels 2 to 5, upward jumps were estimated significantly higher than downward jumps ($t \geq 2.40$, $p \leq 0.011$), based on pairwise one-tailed Welch's t-tests. For level 6, however, the upward jump was significantly lower than the downward jump ($t(29.24) = -1.93$, $p = 0.032$). This may be explained by the extreme reverberant tail of level 6, leading to the result that participants simply did not notice this change.

The instantaneous delay that was computed in the beginning was further analyzed in a similar way

8.3. "RadioReverb": room reverberation as ambient communication channel

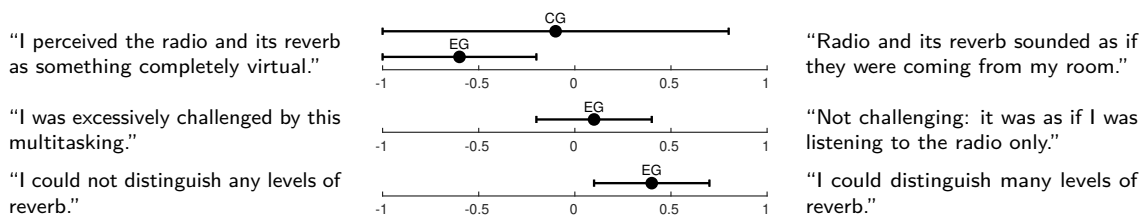


Figure 8.16.: Average rating and standard deviation for the questions on the participants' impressions.

as the estimated reverberation, by computing the average instantaneous delay per time segment and participant. However, when these data were pooled over all participants, no significant effects could be observed.

Finally, Fig. 8.16 shows the numerical results of the questionnaires. EG participants indicated that their listening experience seemed more virtual compared to the CG participants. Mapping of the original range $[1, 7]$ to a range of $[-1, 1]$ produced mean values for CG -0.1 (SD 0.9) and for EG -0.6 (SD 0.4). The EG participants did not feel excessively challenged by listening to the radio while estimating the reverb (mean 0.1 , SD 0.3); however, they found it more challenging than only listening to the radio (CG). Furthermore, they were rather confident of their estimate (mean 0.4 , SD 0.3).

8.3.6. Discussion

The results of the experiment show that listeners were able to discriminate between the seven levels of reverberation with a probability of superiority (i.e., correct answers) of 75%. This probability is often used as a threshold probability when measuring the JND in pairwise comparisons. It shows that our guessed JNDs from the informal listening test were not so far from reality. As this probability implies 25% incorrect identifications, it is usually too low for a practical application in sonification. For an exceptional 97.5% correct answers, only three levels of reverberation can be used. Figure 8.14 allows to obtain the number of discriminable levels (within the fixed parameter range) depending on the probability of superiority that is necessary for a given application.

A sequential bias was found for upward and downward changes of reverberation, leading to a hysteresis curve as shown in Fig. 8.15. This sequential effect is frequently found in such magnitude estimation experiments (see, e.g., Petzschner et al. 2015): The distances between subsequent levels were gen-

erally exaggerated in the participants' estimations. This means that the same true reverberation led to a higher estimated reverberation if the preceding level was lower, and to a lower estimated reverberation if the preceding level was higher. In addition, the absolute estimated magnitudes were compressed compared to the true reverberation; i.e., participants mapped the parameter range to a smaller range. This may be attributed to the fact that the slider did not account for estimations that exceed the true parameter range.

The average instantaneous delay of 1s that was measured, describing the response time of the participants, shows that the sonification system is indeed usable in real-time applications that require user action within physiological reaction times.

We could identify two clusters of participants. The CF participants focused more on the content of the radio features, while the SF participants focused mainly on the sonification and basically neglected the information within the radio features (see Fig. 8.12). To ensure a valid scenario of peripheral sonification, as envisaged for future applications, only CF participants were taken into account for further statistical analysis. While the results may still differ from a true ecologically valid experiment with a more complex and challenging main task, we still assume that the envisaged design of an ecologically valid peripheral listening experiment was sufficiently fulfilled.

According to the post-hoc survey, participants were not excessively challenged by the task and felt confident concerning their estimates. The CG participants who were exposed to the radio features at constant reverberation had less negative feelings about the reverberation than the EG participants who generally regarded it as very unnatural. However, this attitude might be attributed to the situation of listening to a virtual radio set in a virtual room through headphones. We assume that in this specific setting, the sonification was not able to blend with the physical soundscape in the same way

as if it were rendered by hidden loudspeakers rendering the auditory augmentation of the whole soundscape including a physical radio set. In comparison to the previously envisaged laboratory experiment with loudspeakers but in a somehow unnatural and uncomfortable studio, the implementation at home is still considered even more ecologically valid.

8.4. “CardioScope”: an augmented stethoscope using ECG data



This section is based on the original publication which was created in teamwork by Aldana Blanco, Weger, Grautoff, Höldrich, and Hermann (2019). My main contribution was the synchronous data acquisition of ECG and PCG.

CardioScope is the prototype of an augmented stethoscope that is designed to facilitate cardiac diagnosis and monitoring. Usually, a physician listens to the sound of the heart through a stethoscope. This listening practice is called auscultation. The recording of an auscultation of the heart is called phonocardiogram (PCG). In CardioScope, this sound is augmented in real time by information that is derived from the electrocardiogram (ECG), to make interesting aspects of the signal more salient. The ECG signal is usually less noisy than the PCG and thus allows a segmentation of the signal into several time segments of the cardiac cycle (the time between two heartbeats). If the stethoscope is regarded as an auditory magnifying glass to zoom in to a specific spatial region of the heart, CardioScope allows to zoom in to a specific time segment within each cardiac cycle. While the system is only briefly summarized here; a comprehensive description is given by Aldana Blanco et al. (2019).

8.4.1. Electrocardiography and the stethoscope

Figure 8.17 depicts the time signals of ECG and PCG in conjunction with pressures and volumes from cardiac physiology. The heart sounds captured by the stethoscope are physiologically generated by the closing of the heart valves; however, the actual cause are not the valves themselves but the turbulences which result from their closing. During the cardiac cycle, the PCG contains three distinct sound events (labeled 1st, 2nd, and 3rd in Fig. 8.17). The 1st heart sound appears when the mitral valve closes and the aortic valve opens. The 2nd heart sound

appears when the aortic valve closes and the mitral valve opens. Both 1st and 2nd sound therefore involve two separate physiological events (and thus sound events) which are hard to discriminate by ear during auscultation. The 3rd is a very low frequency sound that is produced during ventricular filling. Its actual cause is not yet fully understood and subject to current research (e.g., Omar and Guglin 2017). (Lilly 2016, 26ff)

Additional heart sounds that diverge from the normal sound are called murmurs. They allow a physician to conjecture about abnormal turbulences, e.g., due to a certain valve that is not opening or closing correctly.

In electrocardiography, a graph of voltage over time—the ECG—is created by measuring the electrical activity of the heart through electrodes that are placed on the skin. As depicted in Fig. 8.17, the ECG of a healthy person involves three main components. The P-wave represents the depolarization of the atrium, i.e., the voltage drop in neural activity before reaching the action potential (Colman 2009, p. 201). The QRS-complex with the prominent R-peak which is usually the global maximum of the cardiac cycle represents the depolarization of the ventricles. The T-wave shows the repolarization of the ventricles, i.e., the voltage change back to resting state, including some overshoot, after reaching the action potential. (Lilly 2016, 74ff)

Figure 8.17 and the corresponding physiological mechanisms suggest that activity in the ECG generally precedes activity in the PCG: the 1st and 2nd sound event come from the turbulences that result from muscle activities that are initiated by the nervous action potentials. As the reference points (P, Q, R, S, and T) are generally easier to detect than the sound events (1st, 2nd, 3rd), they are used to label the PCG signal in order to augment segments that are relevant to the physician under certain conditions.

8.4.2. Synchronous acquisition of ECG and PCG

In the envisaged application in a medical context, both ECG and PCG signals are usually provided by a medical data acquisition system. To capture both domains in synchrony, it is beneficial to process them in the same domain. As PCG is anyway already an audio signal—our preferred signal type as sound engineers—we capture the electrical ECG signal in synchrony through the same M-Audio Mobile Pre

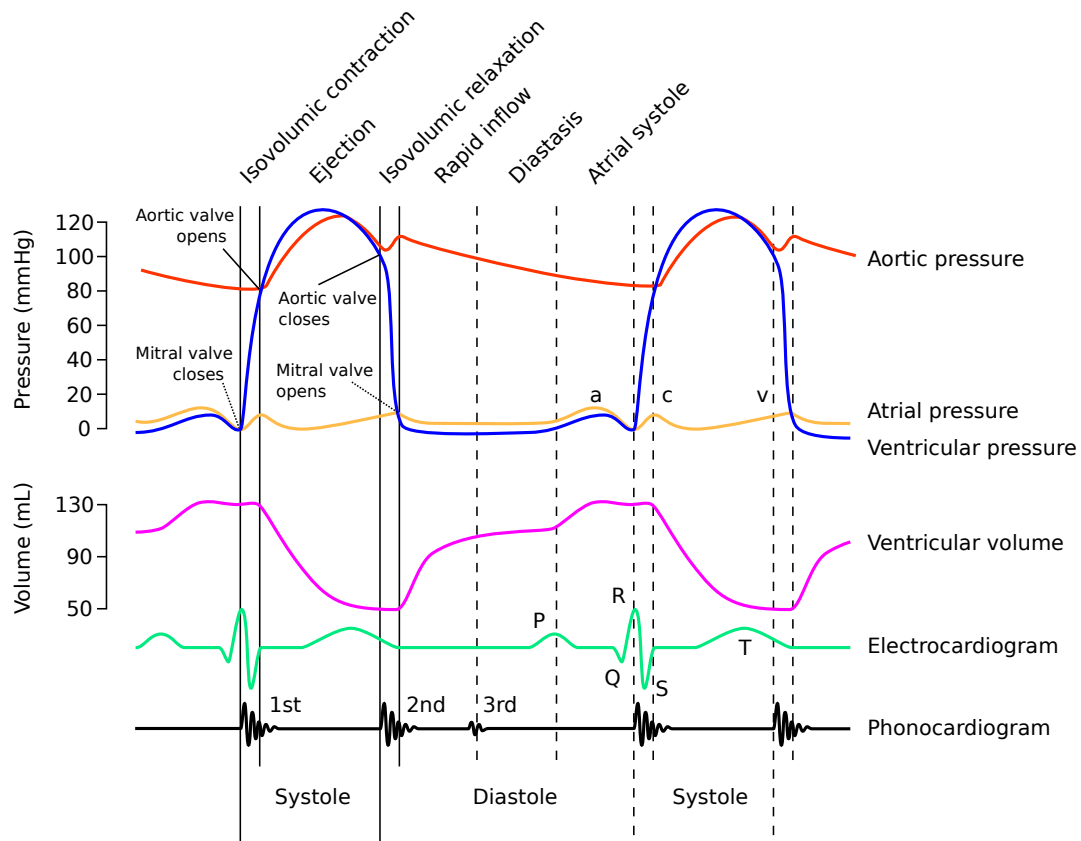


Figure 8.17.: Wiggers diagram. Adapted from Wikimedia Commons (users adh30, DanielChangMD, DestinyQx, and xavax). Licensed under CC BY-SA 4.0.

8. Case studies of auditory augmentations

USB audio interface.

For ECG, we use the ECG Sensor that comes with the BITalino (r)evolution plugged kit⁷ (Plácido da Silva et al. 2014). Instead of connecting it to the BITalino board via its USB connector, we soldered cables to directly connect it to the audio interface. It requires an input voltage V_{cc} between 2.0 and 3.5 V (see Gomes 2020) which we supply through a 3.7 V rechargeable battery. An additionally required $V_{cc}/2$ is provided through a simple voltage divider with two 10 k Ω resistors. The ECG sensor then outputs a voltage between 0 V and V_{cc} . For a stable reference voltage of 3.3 V, we would need an additional voltage regulator. While this would allow the reconstruction of absolute voltage values from the digital signal, we omitted this step in the first prototype. To create a centered and bipolar audio signal, we rely on the analog DC-removal high-pass filter that is already built into the audio interface. The centered signal between $-V_{cc}/2$ and $+V_{cc}/2$ then sufficiently matches the specification of a +4 dBu line level audio signal which reaches from $-V_{peak}$ to $+V_{peak}$, with $V_{peak} = 1.736$ V.

For PCG, i.e., recording of the stethoscope sound, we use a DocCheck Advance II dual head stethoscope chest piece attached to a short rubber tube. An AKG C417 PP miniature microphone is inserted into the open end of the tube and connected to the microphone preamplifier of the audio interface.

Figure 8.18 shows the whole signal acquisition system including ECG sensors and stethoscope. It must be noted that our prototypical procedure involves a direct electrical connection between the ECG sensors and thus electrodes, the audio interface, the computer, and the microphone. Apart from the disregard of any medical safety standards, this setup is prone to noise. While we argue that our system won't bear a greater risk than a microphone, future prototypes should surely involve at least some kind of isolation for the ECG part via optocouplers, to improve the signal-to-noise ratio and reduce electrical risk.

8.4.3. ECG-informed auditory augmentation of heart sounds

In order to facilitate the identification of heart murmurs through auscultation, we propose an auditory augmentation that emphasizes relevant parts and suppresses irrelevant parts of the sound by means of amplitude modulation. Robust information for

⁷BITalino: <https://bitalino.com>

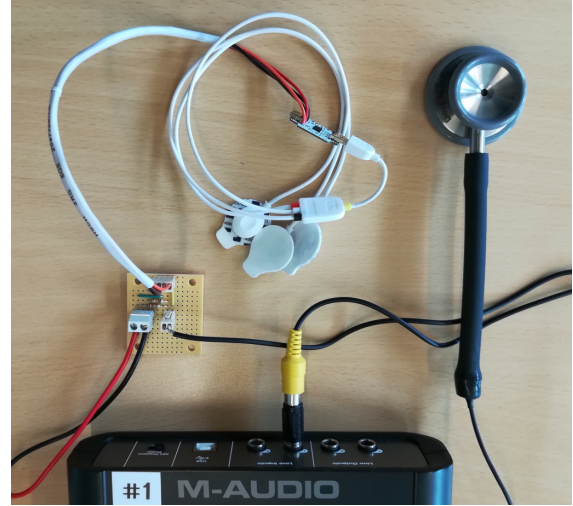


Figure 8.18.: The system for synchronous signal acquisition of ECG and PCG through the same audio interface.

the segmentation of the cardiac cycle is extracted from the ECG in real time by detecting the reference points. The idea of segmenting the PCG based on information from ECG is not new and has already been proposed by others (Malarvili et al. 2003; Giordano and Knaflitz 2019). We propose ECG-synchronized auditory augmentation of heart sounds in real time by means of amplitude modulation.

In our prototype we use only the R-peak which occurs just before the 1st sound event and delivers a robust trigger signal to signify the start of the cardiac cycle. Due to its prominence, it is generally easy to detect by onset detection algorithms. In the prototypical offline implementation, the R-peaks are detected a priori.

Relative to the latest R-peak, we place a raised cosine or Hann window in order to augment a certain position within the cardiac cycle. The pure window itself is defined as

$$W(t) = \sin^2(\pi t) \quad (8.7)$$

for $0 \leq t \leq T_i$. Otherwise $W(t)$ is 0.

From the last detected R-peaks, we predict the duration T_i of the currently running cardiac cycle (starting with the latest R-peak). The start of the window is set relative to the duration T_i of the cardiac cycle by the lag ratio β between 0 and 1. The duration of the window is set by relative length α between 0 and 1. An additional gain g_a blends between the unprocessed signal and the window function. The final equation for the amplitude mod-

ulation signal $e_i(t)$, i.e., the applied envelope, for the given segment i becomes

$$e_i(t) = (1 - g_a) + g_a W\left(\frac{t - \beta T_i}{\alpha T_i}\right). \quad (8.8)$$

Sound examples are provided by Aldana Blanco et al. (2019) in Source 8.4. Sound H1 is the recording of a healthy heart, recorded with the method described in Sec. 8.4.2. In order to test the presented auditory augmentation on heart sounds of non-healthy patients, the large dataset from the “Classifying Heart Sounds Challenge” by Bentley et al. (2012) was used. Sounds P1.1 and P2.1 are the raw recordings of pathological heart sounds which contain different types of murmurs. In sounds P1.2 and P2.2, these murmurs are made more salient by the auditory augmentation, using the parameters $[g_a: 0.9, \beta: 0.4, \alpha: 0.3]$ and $[g_a: 0.9, \beta: 0.5, \alpha: 0.4]$, respectively.

8.4.4. Discussion

As described in detail by Aldana Blanco et al. (2019), the presented auditory augmentation was evaluated by two physicians in a preliminary qualitative test. Overall, their feedback reflects mixed feelings towards the auditory augmentation. While they were generally positive about the benefit of the system, they argued that they would not rely on it for diagnostic purposes and rather stick to classical auscultation in conjunction with other cardiac medical technologies. We attribute this rejection to the fact that they only listened to pre-recorded sounds, completely decoupled from interaction with the patient. We assume that a real-time implementation where physicians were able to tune the parameters of the augmentation (e.g., starting without augmentation, and tuning in only in case of ambiguities) within the interaction loop, the results would be entirely different. This assumption, however, needs to be evaluated in an ecologically valid experiment with a future implementation of the auditory augmentation in compliance with the appropriate medical standards for signal acquisition.

Bibliography

Abrahamse, Wokje et al. (Sept. 2005). “A review of intervention studies aimed at household energy conservation”. In: *Journal of Environmental Psychology* 25.3, pp. 273–291. DOI: 10.1016/j.jenvp.2005.08.002.

Aldana Blanco, Andrea Lorena et al. (2019). “CardioScope: ECG sonification and auditory augmentation of heart sounds to support cardiac diagnostic and monitoring.” In: *Interactive Sonification Workshop (ISon)*. Stockholm, Sweden, pp. 115–122.

Aramaki, Mitsuko et al. (Feb. 2011). “Controlling the Perceived Material in an Impact Sound Synthesizer”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.2, pp. 301–314. DOI: 10.1109/TASL.2010.2047755.

Bentley, Peter et al. (2012). *Classifying Heart Sounds Challenge*. URL: <http://www.peterjbentley.com/heartchallenge/> (visited on 09/22/2021).

Blauert, Jens (1997). *Spatial hearing: the psychophysics of human sound localization*. Rev.ed. MIT press. ISBN: 978-0-262-02413-6.

Bovermann, Till, René Tünnermann, and Thomas Hermann (Apr. 2010). “Auditory Augmentation”. In: *International Journal of Ambient Computing and Intelligence* 2.2, pp. 27–41. DOI: 10.4018/jaci.2010040102.

Bradley, J. S., R. Reich, and S. G. Norcross (Oct. 1, 1999). “A just noticeable difference in C50 for speech”. In: *Applied Acoustics* 58.2, pp. 99–108. DOI: 10.1016/S0003-682X(98)00075-9.

Bradley, Margaret M. and Peter J. Lang (Mar. 1994). “Measuring emotion: The self-assessment manikin and the semantic differential”. In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1, pp. 49–59. DOI: 10.1016/0005-7916(94)90063-9.

Colman, Andrew M. (2009). *Oxford dictionary of psychology*. 3rd ed. New York, NY: Oxford University Press. ISBN: 978-0-19-953406-7.

Cook, Perry R. (2003). *Real Sound Synthesis for Interactive Applications*. 3rd ed. Wellesley, MA: A K Peters. ISBN: 1-56881-168-3.

Cox, T J, W J Davies, and Y W Lam (1993). “The Sensitivity of Listeners to Early Sound Field Changes in Auditoria”. In: *ACUSTICA* 79, pp. 27–41.

Darby, Sarah (2006). *The effectiveness of feedback on energy consumption*. Environmental Change Institute, University of Oxford.

Fischer, Dr Corinna (2007). “Influencing electricity consumption via consumer feedback: a review of experience”. In: *ECEEE Summer Study*, pp. 1873–1884.

Fletcher, Neville H. and Thomas D. Rossing (2010). *The physics of musical instruments*. 2nd ed. New York, NY: Springer. ISBN: 978-1-4419-3120-7.

8. Case studies of auditory augmentations

- Giordano, Noemi and Marco Knaflitz (Apr. 19, 2019). "A Novel Method for Measuring the Timing of Heart Sound Components through Digital Phonocardiography". In: *Sensors* 19.8. DOI: 10.3390/s19081868.
- Gomes, Pedro (2020). *ECG 10082020 - Electrocardiography (ECG) Sensor Data Sheet, Rev. C. PLUX - Wireless Biosignals, S.A.*
- Groß-Vogt, Katharina et al. (June 2018). "Augmentation of an Institute's Kitchen: An Ambient Auditory Display of Electric Power Consumption". In: *International Conference on Auditory Display (ICAD)*. Houghton, Michigan, pp. 105–112. DOI: 10.21785/icad2018.027.
- Groß-Vogt, Katharina et al. (Apr. 5, 2021). "Peripheral Sonification by Means of Virtual Room Acoustics". In: *Computer Music Journal* 44.1, pp. 71–88. DOI: 10.1162/comj_a_00553.
- Hazlewood, William R, Erik Stolterman, and Kay Connelly (2011). "Issues in evaluating ambient displays in the wild: two case studies". In: *CHI*. Vancouver, Canada: ACM, pp. 877–886.
- Iglesia, Daniel (2016). "The Mobility is the Message: the Development and Uses of MobMuPlat". In: *Pure Data Convention*. New York, NY.
- Larsen, Erik et al. (July 1, 2008). "On the minimum audible difference in direct-to-reverberant energy ratio." In: *The Journal of the Acoustical Society of America* 124.1, pp. 450–461. DOI: 10.1121/1.2936368.
- Lilly, Leonard S (2016). *Pathophysiology of Heart Disease*. 6th ed. Wolters Kluwer. ISBN: 978-1-4511-9275-9.
- Malarvili, M.B. et al. (2003). "Heart sound segmentation algorithm based on instantaneous energy of electrocardiogram". In: *Computers in Cardiology*. Thessaloniki Chalkidiki, Greece: IEEE, pp. 327–330. DOI: 10.1109/CIC.2003.1291157.
- McPherson, Andrew P. and Victor Zappi (2015). "An Environment for Submillisecond-Latency Audio and Sensor Processing on BeagleBone Black". In: *AES Convention*. Warsaw, Poland: Audio Engineering Society.
- Miller, George A. (Mar. 1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological Review* 63.2, pp. 81–97. DOI: 10.1037/h0043158.
- Müller-Tomfelde, Christian and Tobias Münch (2001). "Modeling And Sonifying Pen Strokes On Surfaces". In: *Conference on Digital Audio Effects (DAFX)*. Limerick, Ireland.
- Niaounakis, T. I. and William J. Davies (May 15, 2002). "Perception of Reverberation Time in Small Listening Rooms." In: *Journal of the Audio Engineering Society* 50.5, pp. 343–350.
- Omar, Hesham R. and Maya Guglin (Feb. 2017). "Mitral annulus diameter is the main echocardiographic correlate of S3 gallop in acute heart failure". In: *International Journal of Cardiology* 228, pp. 834–836. DOI: 10.1016/j.ijcard.2016.11.254.
- Parthy, Abhaya, Craig Jin, and André van Schaik (2004). "Reverberation for ambient data communication." In: *International Conference on Auditory Display (ICAD)*. Sydney, Australia.
- Petzschner, Frederike H., Stefan Glasauer, and Klaas E. Stephan (May 2015). "A Bayesian perspective on magnitude estimation". In: *Trends in Cognitive Sciences* 19.5, pp. 285–293. DOI: 10.1016/j.tics.2015.03.002.
- Plácido da Silva, Hugo et al. (2014). "BITalino: A Novel Hardware Framework for Physiological Computing." in: *International Conference on Physiological Computing Systems*. Lisbon, Portugal: SCITEPRESS, pp. 246–253. DOI: 10.5220/0004727802460253.
- Schörkhuber, Christian, Markus Zaunschirm, and Robert Höldrich (2018). "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares". In: *DAGA*. Munich.
- Schwartz, Tobias et al. (May 2013). "Uncovering practices of making energy consumption accountable: A phenomenological inquiry". In: *ACM Transactions on Computer-Human Interaction* 20.2, pp. 1–30. DOI: 10.1145/2463579.2463583.
- Seraphim, H.P. (Jan. 1, 1958). "Untersuchungen über die Unterschiedsschwelle exponentiellen Abklingens von Rauschbandimpulsen." In: *Acta Acustica united with Acustica* 8.4, pp. 280–284.
- Soares, Ana Paula et al. (Dec. 2013). "Affective auditory stimuli: Adaptation of the International Affective Digitized Sounds (IADS-2) for European Portuguese". In: *Behavior Research Methods* 45.4, pp. 1168–1181. DOI: 10.3758/s13428-012-0310-1.
- Tillman, J. Donald (2005). *A Discrete FET Guitar Preamp*. URL: <https://web.archive.org/web/20210608211707/http://www.till.com/articles/GuitarPreamp/> (visited on 06/08/2021).
- Tünnermann, René, Jan Hammerschmidt, and Thomas Hermann (2013). "Blended sonification–sonification for casual information interac-

- tion." In: *International Conference on Auditory Display (ICAD)*. Lodz, Poland.
- Warburton, G. B. (June 1954). "The Vibration of Rectangular Plates". In: *Proceedings of the Institution of Mechanical Engineers* 168.1, pp. 371–384. DOI: 10.1243/PIME_PROC_1954_168_040_02.
- Weger, Marian and Robert Höldrich (2019). "A hear-through system for plausible auditory contrast enhancement". In: *Proceedings of Audio Mostly*. Nottingham, UK: ACM, pp. 1–8. DOI: 10.1145/3356590.3356593.
- Yang, Yung-Hao and Su-Ling Yeh (Apr. 2014). "Unmasking the dichoptic mask by sound: spatial congruency matters". In: *Experimental Brain Research* 232.4, pp. 1109–1116. DOI: 10.1007/s00221-014-3820-5.
- Zappi, Victor and Andrew P. McPherson (2014). "Design and use of a hackable digital instrument." In: *Live Interfaces*.
- Zaunschirm, Markus, Christian Schörkhuber, and Robert Höldrich (June 1, 2018). "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint." In: *The Journal of the Acoustical Society of America* 143.6, pp. 3616–3627. DOI: 10.1121/1.5040489.
- Zotter, Franz and Matthias Frank (2019). *Ambisonics - A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Open. ISBN: 978-3-030-17206-0.

9. Auditory contrast enhancement (ACE)

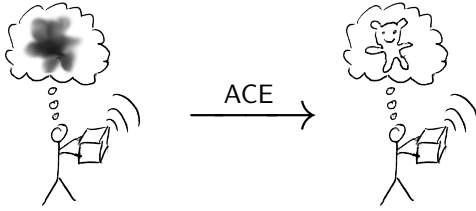


Figure 9.1.: Someone shaking a box to guess its contents from the resulting sound. A task we want to facilitate.



This chapter is based on the following publications: Weger, Hermann, and Höldrich (2019), Weger and Höldrich (2019).

In Chapter 1, we introduced the concept of augmented auditory feedback and auditory augmentation. One main objective of auditory augmentation might be to enhance relevant information that is contained in the sound, so that it becomes more salient to the listener, e.g., by improving the signal-to-noise ratio (SNR). This we call auditory contrast enhancement (ACE).

What might be relevant to users, however, depends on their individual activities, as well as on the type and origin of the observed sound. The general concept is illustrated in Fig. 9.1. We expect high potential for ACE where listening is part of a knowledge-making process. Especially when, for example, scientists, engineers, or physicians rely on their ears during professional routines.

A very common listening practice is to knock on an object to obtain information on its inner structure. This can be seen as a rudimentary form of an impulse response measurement. For example, craftspersons locate cavities and studs behind a wall by thumping in order to find an appropriate spot for a drill hole. Farmers knock on a melon to tell if it is ripe. Violin makers tune the top and back plate of the instrument by holding it lightly between two fingers and tapping it, which reveals the so-called *tap tones*. Similarly, the filling level of a barrel is estimated by ear via knocking. This active listening practice inspired the son of an innkeeper and later physician Leopold Auenbrugger to his “*inventum novum*” (see Bishop 1961; Nuland 2005): *percus-*

sion. As a method of clinical examination, it has become indispensable in the repertoire of modern physicians—together with its passive counterpart, *auscultation*, which is usually performed by using a stethoscope. The stethoscope enhances auditory contrast not only by efficient guidance of the structure-borne sound to the user’s ears, but also by amplification of frequency ranges which are of special interest to the user (Ertel et al. 1971). Despite the medical definitions, percussion and auscultation are here interpreted in a broader sense where the examined object may be any physical object.

We differentiate between two types of auditory contrast. *Intra-stimulus contrast* describes the prominence of those features that characterize a single sound. As an example, for some Stradivarius and Guarnerius violins, the tap tone of the top plate exhibits a prominent resonance between one and two semitones above the middle *C* (Hutchins 1962). For this feature, intra-stimulus contrast can be quantified by the amplitude difference between peak and average of the spectrum. By *inter-stimulus contrast*, we mean the perceptual differences between two or more sounds or groups of sounds. Sticking to the previous example, the tap tone of the back plate is usually one or two semitones higher (Hutchins 1962). Inter-stimulus contrast between the tap tones of top and back plate could be quantified by their spectral differences, i.e., differences in amplitude, frequency, and bandwidth. Apart from this example, both types of auditory contrast can be regarded from different perspectives, e.g., focusing on spectral, temporal, or spectro-temporal features.

ACE therefore comes in two flavors. *Intra-stimulus ACE* is a method to make intrinsic features of a sound more prominent, in order to facilitate their perception and thus improve the conveyance of the underlying information. This is possible in real time. The underlying information may be, for example, physical properties such as material, shape, or size of a cavity. The outcome can be interpreted as a *cartoonification* of the original sound. *Inter-stimulus ACE* is a method for the auditory display of differences between groups of sounds, based on supervised or unsupervised learning from pre-recorded samples (Hermann and Weger 2019). This requires

9. Auditory contrast enhancement (ACE)

the acquisition of training data. However, a thus produced spectral ACE filter can then be applied on an unknown audio signal in real time, to facilitate a classification by ear.

Our goal is to enhance the perception of those sound properties that characterize a sound, while maintaining its original gestalt as good as possible. We assume that this compromise can be achieved by attenuating non-characteristic aspects of the signal, thus leading to reduced spectral, temporal, and informational masking. In the extreme case, a very strong contrast enhancement leads to a cartoonification of the sound, reducing it to only a few very prominent sound attributes. This is conceptually similar to the visual domain where contrast is usually understood as the degree to which areas of an image differ in appearance.

Assuming that a sound is characterized by its unique spectral and temporal structure, an enhancement of this structure may automatically enhance the contrast to other sounds which exhibit a different structure. If, however, two sounds share the same strong characteristics with only minor differences, intra-stimulus ACE could even suppress those differences, leading to reduced inter-stimulus contrast between both. Such “similarity enhancement” might be useful when searching for similarities between stimuli. Otherwise, inter-stimulus contrast enhancement would be the recommended choice (see Hermann and Weger 2019).

Intra-stimulus ACE tries to improve absolute identification of a single sound, while inter-stimulus ACE tries to facilitate discrimination between multiple sounds. Nevertheless, both are tightly connected. In case of the tap tones, the broadband resonances of top and bottom plate differ only slightly in frequency. A bandwidth reduction, as achieved by intra-stimulus ACE, transforms them into more tonal signals with higher pitch strength, and consequently higher inter-stimulus contrast. This effect is schematically depicted in Fig. 9.2.

In summary, we identify two activities which intra-stimulus ACE should improve: (1) identify the physical sound source, as visualized in Fig. 9.1, and (2) discriminate between sounds that are different to each other.

The remainder of this chapter is structured as follows. In Sec. 9.1 we derive an algorithm for real-time intra-stimulus ACE.

The plausibility of the resulting auditory feedback was evaluated in an auditory-visual experiment (Sec. 9.2). The main objective was to find a compromise between high auditory contrast and acceptable

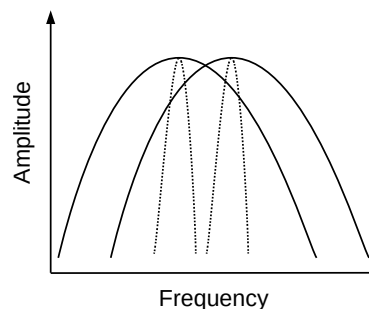


Figure 9.2.: The principle of enhanced inter-stimulus contrast by enhanced intra-stimulus contrast.

plausibility for observed interactions.

The described algorithm for intra-stimulus ACE is supposed to maintain high plausibility of the auditory feedback, as it is based on the same mechanism that is used in the neural networks of auditory nerves and auditory cortex (Kral and Majernik 1996; Pantev et al. 2004), namely lateral inhibition. It describes the inhibitory effect of stimulation in neighboring channels. We further use a physically meaningful decay prolongation with frequency-dependent exponential decay, in order to provide listeners with more time to perceive characteristic partials.

In Sec. 9.3, we present a practical implementation of intra-stimulus ACE based on a microphone-earphone combination. Used as a tool, it works similar to a hearing aid, with the goal to enhance the mentioned listening practices, with a focus on percussion. In other words, the system should help to identify or categorize short impact sounds, as well as to discriminate between them, by ear. Interpretation and decision-making is left to the user.

The sound examples for ACE that are referenced in the text are found in Source 9.1.



9.1. Real-time auditory contrast enhancement

The main applications that are envisaged for real-time ACE are percussion and auscultation — not so much for medical purposes but more for material testing by ear and auditory observation of mechanical processes such as machines. The targeted sounds therefore include transient interaction sounds and environmental sounds, but not speech or music. The focus on real-time application on auditory feedback makes a low-latency implementation necessary. Furthermore, the sounds resulting from

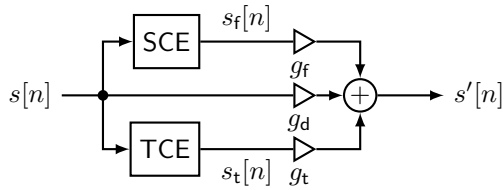


Figure 9.3.: Overall block diagram of real-time ACE.

ACE should maintain some degree of naturalness—they should stay within the limits of plausibility with reference to their individual context and the performed action. While development is performed in Matlab, the real-time algorithm is implemented in SuperCollider.

Figure 9.3 shows the overall block diagram. Output $s'[n]$ is a mix of three signals: (1) the dry input signal $s[n]$ (e.g., coming from a microphone), (2) the output $s_f[n]$ of spectral contrast enhancement (SCE, see Sec. 9.1.1), and (3) the output $s_t[n]$ of temporal contrast enhancement (TCE, see Sec. 9.1.2). Their individual gains are parameterized by two linear cross-fades: (1) between $s_f[n]$ and $s_t[n]$ to intuitively tune to the signal dimension of interest, and (2) between this weighted sum and the original signal (wet and dry) for overall strength of the effect.

9.1.1. Spectral contrast enhancement

Yang et al. (2003) define spectral contrast as “the decibel difference between peaks and valleys in the [magnitude] spectrum”. They describe several algorithms for spectral contrast enhancement, aiming at two applications: (1) compensation of reduced frequency selectivity in hearing-impaired people, and (2) speech enhancement in noise. One of the easiest methods is to exponentiate the magnitude spectrum by a variable exponent, followed by normalization (Boers 1980). This results in a spectral dynamics expansion with respect to the global maximum. Other approaches use linear prediction which works well for speech enhancement where detailed information on the sound source is available (Yang et al. 2003).

A large group of algorithms is based on an analog circuit proposed by Stone and Moore (1992). In principle, the signal is split into a number of frequency bands which are separately processed by a variable gain amplifier and then summed. The gain of each channel is a weighted sum of its own envelope and the envelopes of four neighboring channels; the latter with negative weights. This weighting is

similar to a transversal FIR filter. As result, spectral peaks are amplified while troughs are attenuated. The digital implementation of this algorithm—Yang et al. refer to it as “Cambridge’s method”—works as follows (Yang et al. 2003; Baer et al. 1993):

1. Computation of the spectrum X_k of a (windowed) signal block via FFT, with frequency index k .
2. Calculation of excitation pattern P_k —“the representation of a spectral shape in the auditory system” (Stone and Moore 1992). It resembles a smoothed version of the magnitude spectrum $|X_k|$.
3. The enhancement function E_k is the convolution of P_k with a difference-of-Gaussians (DoG) function. This is similar to a smoothed 2nd derivative. The DoG function is the sum of a positive Gaussian and a negative Gaussian with larger (here: $2\times$) bandwidth. Convolution runs on a scale which quantifies the number of equivalent rectangular bandwidths (ERB) that fit below a certain frequency—the ERB-rate scale (Moore and Glasberg 1996).
4. The enhanced magnitude spectrum $|Y_k|$ is then

$$|Y_k| = P_k \cdot (|E_k| + 1)^{\text{sgn}(E_k) \cdot \rho}, \quad (9.1)$$

where $\rho \geq 0$ controls the strength of the effect.

5. Inverse FFT of $|Y_k|$ combined with the original phase values.

While Cambridge’s method did not improve speech intelligibility—neither analog nor digital—its high potential in “technical” enhancement of spectral contrast, i.e., increasing differences between peaks and valleys, is evident.

Our auditory system achieves spectral contrast enhancement similar to Cambridge’s method. The underlying mechanism is based on Lateral Inhibition (LI) in the neural networks of the auditory nerves and the auditory cortex (Kral and Majernik 1996; Pantev et al. 2004). In general, this process can be described as “the suppression of nervous activity at one place in a receptor field as a consequence of the stimulation of adjacent places in this field” (Houtgast 1972). Besides, for instance, the retina and the skin, such receptor fields are also found along the basilar membrane (Coren et al. 1988; Békésy 1962). Kral and Majernik (1996) used an artificial neural network to model the effect of spectral contrast enhancement in the auditory system via lateral inhibition. Among their simulated scenarios, three extreme cases are of particular interest. (1) Partly

9. Auditory contrast enhancement (ACE)

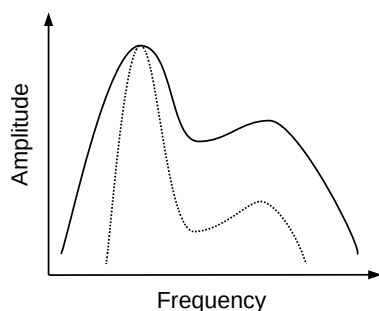


Figure 9.4.: Principle of dynamics expansion.

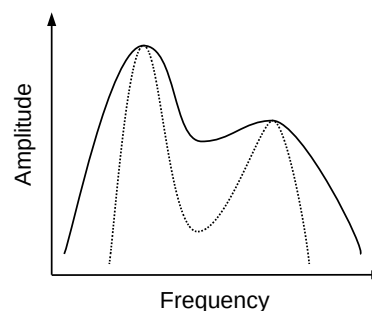


Figure 9.5.: Principle of lateral inhibition.

overlapping band-limited noise signals are narrowed in bandwidth and thus separated. (2) Uniform white noise is effectively suppressed. (3) Uniform white noise where a specific frequency range has been suppressed leads to spikes at the edges of the stop-band—the so-called edge effect.

It seems that in general there are two types of spectral contrast: (1) exponentiation relative to the global maximum (we refer to it as spectral dynamics expansion), and (2) lateral inhibition (we refer to it as spectral sharpening). Figures 9.4 and 9.5 illustrate their effect on a certain spectrum, respectively. It might be interesting to compare these to the visual domain. Spectral dynamics expansion compares to visual contrast control as shown in Fig. 9.6b, while spectral sharpening is actually edge detection (see Fig. 9.6c; the image shows the inverted result)—remember the edge effect demonstrated by Kral and Majernik (1996). In order to achieve something close to cartoonification, as exaggeratedly illustrated in Fig. 9.6d, we would need a combination of both types of contrast. In vision, this would be an overlay of Fig. 9.6b and c, e.g., by multiplying or taking the minimum of both images). In the auditory domain, we would take the maximum of both output spectra. The above considerations suggest that both types of spectral contrast enhancement are necessary, depending on the sound characteristics of interest, and therefore need to be implemented for parallel or serial use.

As we target low latency and real-time operation, the use of FFT—the basis for the majority of speech enhancement algorithms—is not possible. For that reason, frequency separation must be achieved by a filterbank, similar to the analog cir-

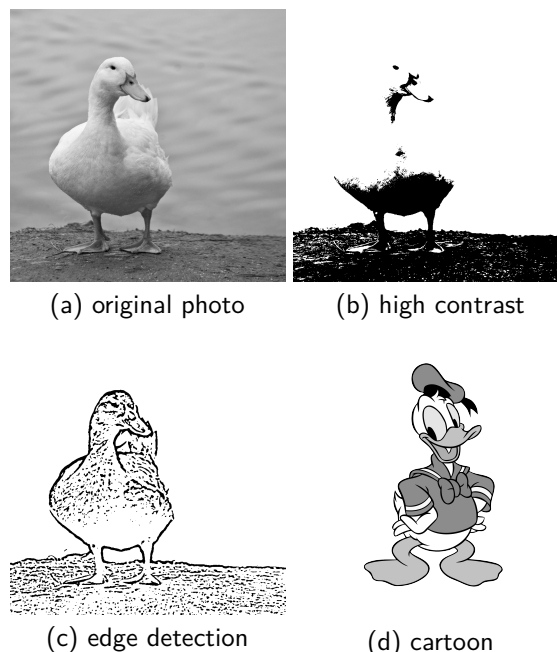


Figure 9.6.: The photo of a white duck in three versions, and the drawing of a famous cartoon duck.¹

cuit by Stone and Moore (1992). We are therefore restricted to operate on a very limited number of frequency bands. Note, however, that Cambridge's method returns an altered version of the excitation patterns—a signal with significantly reduced frequency resolution. An adequate approximation of the excitation pattern can be obtained by a gamma-tone filterbank (GTFB)—a widely used model for the auditory filters (Patterson et al. 1987). If the filters' center frequencies are equally spaced on the ERB-rate scale (and set to constant bandwidth in parts of the ERB), they simulate an equal spacing on the basilar membrane. The lower bands exhibit a smaller bandwidth in Hz, leading to longer impulse response and group delay. This implies a trade-off

¹Fig. 9.6a–c: Anne Davis, <http://flickr.com/anned/>, Creative Commons Attribution NonCommercial (CC BY-NC) 2.0 Generic License. Fig. 9.6d: <http://pngimg.com/>, CC BY-NC 4.0 International License.

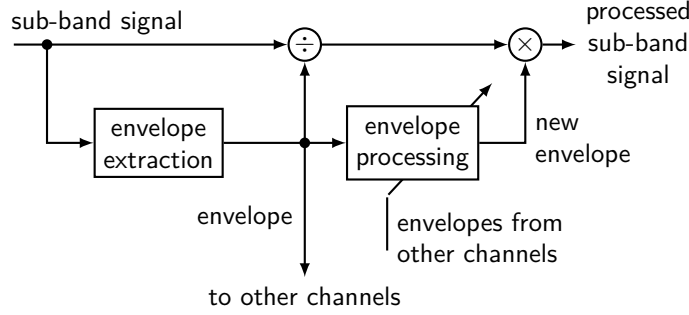


Figure 9.7.: Simplified block diagram of one channel of sub-band processing.

between frequency resolution and group delay towards low frequencies, which needs to be taken care of.

The excitation pattern is expressed by the energy distribution across sub-bands, calculated via their channel envelopes. Depending on the implementation of the gammatone filter, it can also output the imaginary part of the resulting signal, in addition to the real output. An accurate estimation of the signal envelope is then given by the magnitude of the complex filter output. A suitable implementation is the one by Hohmann (2002), which is available for Matlab², Pd³ and SuperCollider⁴; in the latter case, a small modification of the source code is needed in order to return the imaginary part. We use 60 4th-order filters with center frequencies from 50 Hz to 20 kHz, overlapping at their -4 dB cutoff frequency (as used by Noisternig 2017, p. 74). During resynthesis, i.e., summation of the processed sub-bands, their different group delays are usually compensated by individual time-delays, in order to reduce ripple in the output spectrum. We circumvent such additional latency by weighting the sub-bands with alternating signs, as proposed by Noisternig (2017, pp. 72–73).

The overall block diagram of the proposed algorithm for spectral ACE in Fig. 9.8 illustrates the general idea described above. In summary, the input signal $s[n]$ is split into K sub-bands $c_k[n]$ by a gammatone filterbank with K channels; k is the channel index. The actual spectral contrast enhancement

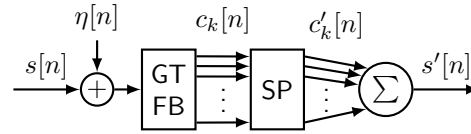


Figure 9.8.: Overall block diagram of spectral ACE

is done within the sub-band processing block (SP). The sum of the processed (real-valued) sub-bands $c'_k[n]$ then forms the enhanced output signal. Within SP, all channels are treated equally, as shown in Fig. 9.7. While the gammatone filterbank accounts for the $1/f$ proportionality of signal energy, this might not be enough for many natural signals which may exhibit even stronger high-frequency loss. This could lead to overly damped high-frequency content in the output. This effect is reduced by a pair of shelving filters—one boosting high frequencies of $s[n]$ before feeding it to the gammatone filterbank, and another one inverting the effect of the first one by attenuation after resynthesis/summation.

Each channel $c_k[n]$ individually passes sub-band processing as shown in Fig. 9.9. First, the sub-band envelope $e_k[n]$ is extracted by taking the absolute value of the complex signal $c_k[n]$. This envelope then successively passes three stages: lateral inhibition (LI, see Sec. 9.1.1.1), exponentiation (EX, Sec. 9.1.1.2), and decay prolongation (DP, Sec. 9.1.1.3). The processed envelope $e'_k[n]$ is finally applied to the real part of the sub-band signal $c_k[n]$ by multiplication with the ratio between processed and original envelope (see Eq. 9.2). Both envelopes are low-pass filtered by a leaky integrator with time-constant $\tau = 2$ ms to suppress disturbing artifacts which occur at high amplitude ratios, especially at low overall volume. For regularization, a small value $\delta = 10^{-5}$ is added to the denominator (assuming audio signals in the range between -1

²Matlab implementation of the used gammatone filterbank (Hohmann 2002):

http://medi.uni-oldenburg.de/download/demo/gammatone-filterbank/gammatone_filterbank-1.1.zip

³Audition library for Pure Data:

<http://lumiere.ens.fr/Audition/tools/realtime/>

⁴AuditoryModeling UGens from SC3 Plugins:

<https://github.com/supercollider/sc3-plugins>

9. Auditory contrast enhancement (ACE)

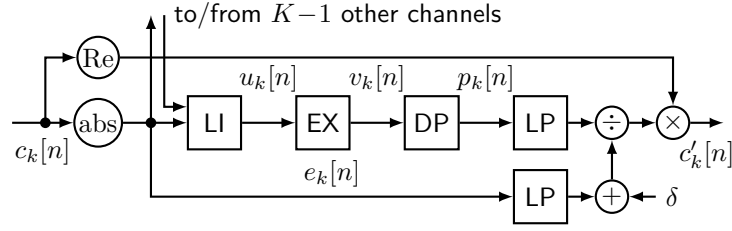


Figure 9.9.: Detailed block diagram of one channel of sub-band processing (SP)

and 1).

$$c'_k[n] = \text{Re}\{c_k[n]\} \cdot \frac{e'_k[n]}{e_k[n] + \delta}. \quad (9.2)$$

9.1.1.1. Spectral sharpening by lateral inhibition

One problem we see in Cambridge's method (Eq. 9.1) is that it not only dampens spectral valleys but also amplifies spectral peaks. This uncontrolled amplification of the signal can be avoided by restricting the enhancement function E_k to negative values.

We first define an inhibition term $T_k[n]$ which quantifies the overall energy in the neighboring sub-bands. If it is larger than the energy in the observed band, then this band is attenuated. Calculation of the inhibition term is based on the sub-band envelopes $e_k[n]$ that are low-pass-filtered by a leaky integrator (with time constant τ_{LI}), which leads to $\tilde{e}_k[n]$. The resulting slow attack time suppresses inhibition caused by short spikes in neighboring bands, while the decay adds an aftereffect to the lateral inhibition.

We base the calculation of the neighboring bands' weights on the DoG function as in Cambridge's method. The ratio between the bandwidths of the two Gaussians controls the sharpness of the resulting spikes in the spectrum. As our approach anyway restricts sharpening to the bandwidths of the used filters (which is quite *unsharp*), we reduce the positive Gaussian to a minimum, being a Dirac delta impulse. This way, extreme enhancement (large ρ) would inhibit all frequency bands except those which describe local maxima. The bandwidth of the negative Gaussian is set via its standard deviation σ in ERB rate.

For the lowest and highest sub-band, neighbors of significant weight are outside the scope of the filterbank. A zero-padding (insertion of zero-valued virtual bands on both sides) would introduce an

unwanted edge-effect at the lowest and highest sub-band ($k=1$ and $k=K$, respectively), similar to the simulation by Kral and Majernik (1996). Therefore, two virtual sub-bands (copies of sub-bands 2 and $K-1$) are introduced as sub-bands 0 and $K+1$, respectively (copying the edge bands themselves would half a potential contrast in those bands). The inhibition term $T_k[n]$ then becomes

$$T_k[n] = \sqrt{\frac{1}{2\gamma_k^-} \sum_{i=0}^{k-1} \gamma_{i,k} \cdot \tilde{e}_i^2[n] + \frac{1}{2\gamma_k^+} \sum_{i=k+1}^{K+1} \gamma_{i,k} \cdot \tilde{e}_i^2[n]}, \quad (9.3)$$

where $\gamma_{i,k}$ is a Gaussian function, with center frequencies f_c of the filters given in ERB rate:

$$\gamma_{i,k} = \exp\left(-\frac{(f_{c,i} - f_{c,k})^2}{2\sigma^2}\right). \quad (9.4)$$

The scaling factor can be omitted, as the weights are anyway normalized for the lower and upper neighbors individually:

$$\gamma_k^- = \sum_{i=0}^{k-1} \gamma_{i,k} \quad \text{and} \quad \gamma_k^+ = \sum_{i=k+1}^{K+1} \gamma_{i,k}. \quad (9.5)$$

This scaling ensures that a signal with equal envelopes, i.e., in which $e_k[n]$ is the same for all k , implies $T_k[n] = \tilde{e}_k[n]$, and therefore leads to unchanged envelopes. Due to the ERB-scaled gammatone filterbank, this is the case for a pink noise signal which exhibits a magnitude spectrum that is proportional to $1/f$. This relation approximates the decrease in energy towards high frequencies, that is common to many natural sounds. In analogy to Eq. 9.1, the sharpened envelopes $u_k[n]$ then become

$$u_k[n] = e_k[n] \cdot \min\left\{\left(\frac{\tilde{e}_k[n]}{T_k[n]}\right)^\rho, 1\right\} \quad (9.6)$$

The amount of spectral sharpening is set by the parameter $\rho \geq 0$. As the quotient $\tilde{e}_k[n]/T_k[n]$ is restricted to values below 1, any $\rho > 0$ literally suppresses lower quotients.

The effect of spectral sharpening is demonstrated by knocking with knuckles on a wooden plate. Listen to the signal without and with spectral ACE (sounds S1.1 and S1.2, respectively, in Source 9.1). Corresponding spectrograms are shown in Fig. 9.10a–b. Parameters have been set to values which work well for most signals: $\rho = 30$, $\sigma = 3$ ERB, and smoothing with $\tau = 7$ ms. It is apparent that the described algorithm effectively suppresses spectral troughs while leaving local maxima as narrow-band regions with their original amplitude. In addition, the broadband background noise is reduced to some high-frequency artifacts of the recording which are now clearly audible. A ρ larger than 30 does not seem to bring any benefit for spectral sharpening; the signal is already reduced to its local maxima. Additional contrast is achieved by spectral dynamics expansion, as explained in the next section.

9.1.1.2. Spectral dynamics expansion by exponentiation

The goal of spectral dynamics expansion is to attenuate frequency bands with low energy while pulling those with high energy, above a certain threshold value, up to the running global maximum. In contrast to spectral sharpening, this approach should not attenuate broadband regions in the spectrum if they are prominent enough. On the downside, it will suppress even very prominent local maxima if they appear below the threshold.

Spectral dynamics processing is achieved by exponentiation of the magnitude spectrum — inspired by the simple algorithm originally proposed by Boers (1980). In our case, each envelope $u_k[n]$ is scaled with respect to the global maximum of all (smoothed) envelopes (see Eq. 9.7). As gain factor, we use the quotient of the smoothed envelope $\tilde{u}_k[n]$ and a fraction of the instantaneous maximum of all smoothed envelopes ($\mu\tilde{u}_{\max}[n]$). The exponent $\beta \geq 0$ sets the amount of expansion; $0 < \mu \leq 1$ is the relative threshold. Gain is clipped at $\tilde{u}_{\max}[n]/\tilde{u}_k[n]$ so that $u_k[n]$ does not exceed the maximum of all sub-band envelopes.

$$v_k[n] = u_k[n] \cdot \min \left\{ \left(\frac{\tilde{u}_k[n]}{\mu\tilde{u}_{\max}[n]} \right)^\beta, \frac{\tilde{u}_{\max}[n]}{\tilde{u}_k[n]} \right\} \quad (9.7)$$

with the (instantaneous) global maximum

$$\tilde{u}_{\max}[n] = \max_k \{ \tilde{u}_k[n] \}. \quad (9.8)$$

Listen again to the enhanced signal from the previous section (Snd. S1.2 / Fig. 9.10b). Additional contrast is achieved by feeding this signal into spectral dynamics expansion (Snd. S1.3 / Fig. 9.10c). Furthermore, the background noise is gone. The parameters have been set to $\mu = 0.8$ and an extreme value of $\beta = 8$, leading to a spectral gate where values below $\mu\tilde{u}_{\max}[n]$ are almost completely suppressed while values above approach the global maximum.

Contrary to spectral sharpening, spectral dynamics expansion can also be used to exaggerate broadband regions in the spectrum. This is demonstrated in Snd. S2.1 and S2.2 with the recording of a vintage printing machine, with noise from a pneumatic system.

9.1.1.3. Decay prolongation by envelope processing

Spectral resolution and pitch impression takes time. What if we gave listeners more time to perceive a sound by prolonging it through artificial decay? Looking just at the envelope of the signal, the desired effect is depicted in Fig. 9.11. Such an effect could be achieved in a natural way via reverberation. Dombois and Eckel argued that reverberation might even be used to enhance audifications, as it may facilitate discrimination between short transient sounds (Hermann et al. 2011, p. 315). Kourmura and Furukawa (2017) showed that reverberation deteriorates material identification, at least for a while, until listeners adapt to the reverberation. This adaptation, however, is not transferable between speech and impact sounds, and must be learned individually. Such natural reverberation, of course, is not correlated to the stimulus itself, but just convolves it with an arbitrary impulse response. A completely “transparent” reverberation whose impulse response has a white magnitude spectrum might already lead to better results.

Yet another problem is the broadband spectrum of the transient sounds — any artificial reverberation will therefore mask succeeding parts completely with broadband noise. Even if the resonances are sharpened through spectral contrast enhancement as derived in Sec. 9.1.1, a short transient signal in a single sub-band still results in a broadband signal at the output. However, if artificial decay is applied to

9. Auditory contrast enhancement (ACE)

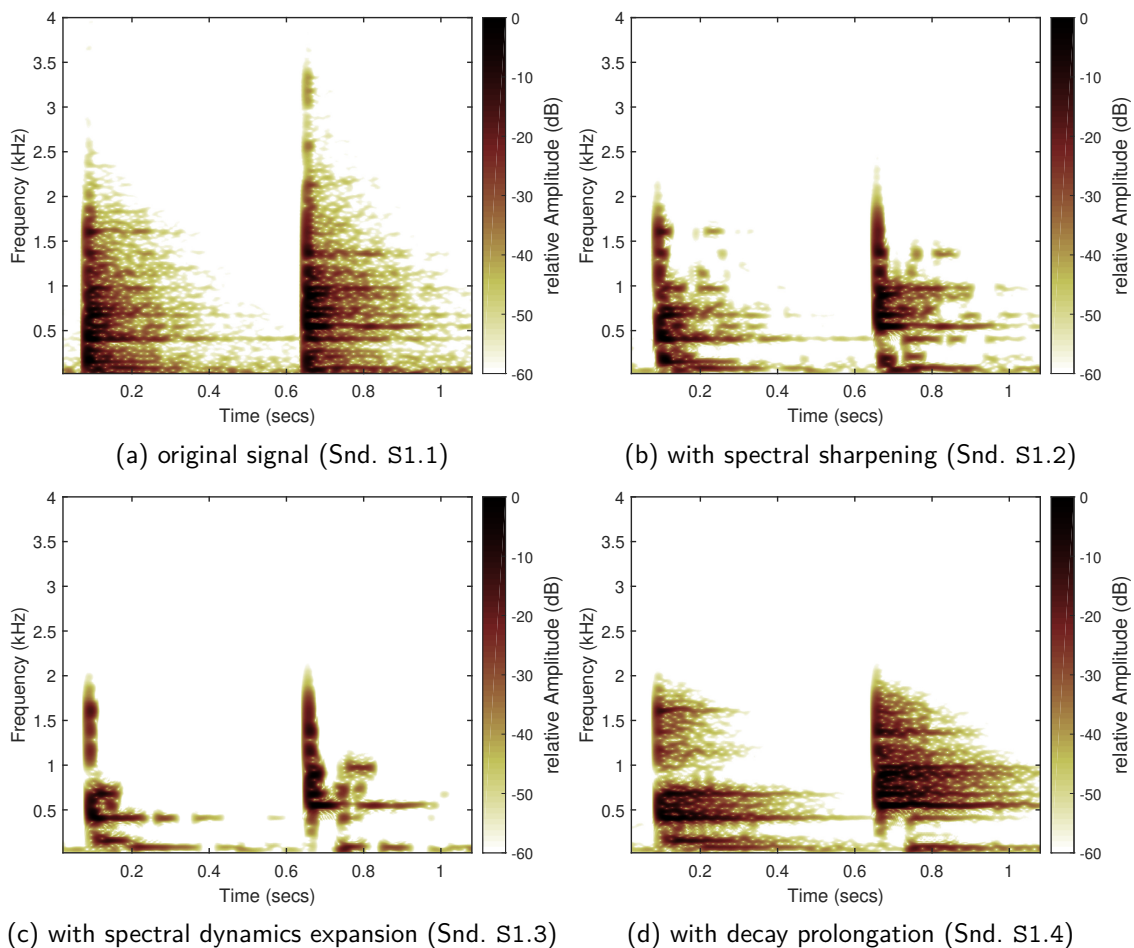


Figure 9.10.: Spectrograms of a test sound in 4 conditions: (a) original recording, (b) with spectral sharpening, (c) with spectral sharpening and spectral dynamics expansion, (d) with spectral sharpening, spectral dynamics expansion, and decay prolongation. The sound files are available in Source 9.1.

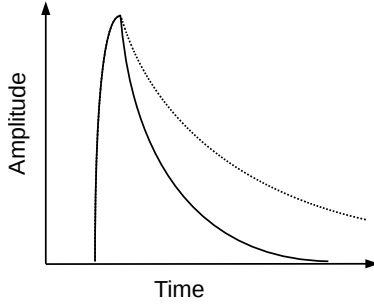


Figure 9.11.: Principle of decay prolongation.

the individual sub-band envelopes, their bandwidths are reduced and more time is given to the listener to gain a pitch impression. The enhanced sub-band envelopes after lateral inhibition and exponentiation may still contain short spikes which are not visible in the spectrogram of Fig. 9.10b–c, but which would have a huge impact if the sub-band envelopes were decayed as they are. Therefore, the envelopes must be smoothed before decay prolongation. As this further smears the envelopes in time, we instead split them into a transient part and a decay part. Only the decay part receives decay prolongation; both are re-combined afterwards.

We first introduce two simple non-linear low-pass filters based on a leaky integrator. env_a has a smooth attack but instant decay, while env_d has a smooth decay but instant attack. env_a is given in Eq. 9.9 for an arbitrary input signal $x[n]$ and output signal $y[n]$. env_d follows the same equation, but with flipped direction of the inequality sign, leading to a naturally-sounding exponential decay.

$$y[n] = \begin{cases} |x[n]|, & \text{for } |x[n]| \geq y[n-1] \\ (1-\alpha)|x[n]| + \alpha y[n-1], & \text{otherwise} \end{cases} \quad (9.9)$$

The amount of smoothing is set via the smoothing factor α . A more convenient parameterization can be achieved via time constant τ or -60 dB reverberation time T_{60} :

$$\alpha = \exp\left(-\frac{1}{\tau F_s}\right) = \exp\left(-\frac{\ln(1000)}{T_{60} F_s}\right) \quad (9.10)$$

where F_s is the sampling frequency.

The envelope with smoothed attack $\text{env}_a\{v_k[n]\}$ is fed to decay prolongation, while the residuum ($v_k[n] - \text{env}_a\{v_k[n]\}$) containing only the attack part is added back to the result, leading to the

output signal of decay prolongation $p_k[n]$:

$$p_k[n] = \text{env}_d\left\{\text{env}_a\{v_k[n]\}\right\} + v_k[n] - \text{env}_a\{v_k[n]\} \quad (9.11)$$

Due to the normalization with the original envelopes (Eq. 9.2) the decay is fed by intrinsic signal components of the sub-band signals in the relevant frequency region. In order to supply sufficient signal energy in the case of large SNR combined with long decay prolongation, a pink noise signal $\eta[n]$ is added to the input signal just before feeding it to the gammatone filterbank (see block diagram in Fig. 9.8a); at a level below the threshold of hearing, but enough to synthesize literally infinite decay. As internal signal processing on any eligible platform offers at least 32 bit floating-point precision, a noise level of around -96 dBFS is more than enough.

A constant decay time over the whole frequency range leads to an unnatural amplification of high frequencies, as damping usually increases with frequency. We chose a rough approximation by setting T_{60} inversely proportional to the center frequency, but clipped below 1 kHz.

Sound example S1.3 from Source 9.1 and Fig. 9.10d show the effect of decay prolongation on the enhanced signal from Sec. 9.1.1.2 (Snd. S1.3 and Fig. 9.10c). For this example, reverberation time T_{60} at 1 kHz was set to 0.5 s. The time constant for transient separation was set to 7 ms. It is clearly visible and audible that relevant partials are significantly extended in time.

9.1.1.4. Vibrato expansion by frequency shifting

At this point, the question arises if also vibrato effects, i.e., frequency modulation, can be exaggerated. The goal would be to enlarge the frequency range of any existing vibrato in the signal, as well as to exaggerate chirps (short tones that rise or fall in frequency).

The instantaneous frequency $f_k[n]$ for each sub-band can be calculated directly from the complex output $c_k[n]$ of the gammatone filter via differentiation of the unwrapped instantaneous phase. However, the computation via zero-crossing rate led to more stable results. An estimate of the instantaneous frequency $f_k[n]$ is obtained by the halved zero-crossing rate which is clipped at the lower and upper -4 dB cutoff frequency of the filter (the frequency at which the next sub-band should take over) and then smoothed with a 2nd-order low-pass filter.

9. Auditory contrast enhancement (ACE)

A quasi-stationary frequency $\tilde{f}_k[n]$ is obtained by low-pass filtering at a lower cutoff frequency. We chose a cutoff frequency of 10 Hz for the instantaneous frequency and 1 Hz for the quasi-stationary frequency. The difference between those two then leads to an estimate of the instantaneous frequency modulation. The modulation depth can then be increased through a frequency shift by the weighted frequency difference $\Delta f_k[n]$:

$$\Delta f_k[n] = \lambda \cdot (f_k[n] - \tilde{f}_k[n]) \quad (9.12)$$

where $\lambda \geq 0$ is a multiplicative factor which controls the strengths of the effect.

The frequency difference $\Delta f_k[n]$ is then applied via single-sideband modulation (SSB). The linear frequency shift will, of course, strongly distort the harmonic relationship between partials in the signal. Each sub-band signal is frequency-shifted by the individual frequency difference:

$$d_k[n] = \text{Re}\{c_k[n]\} \cos(\Delta\omega_k[n] \cdot n) - \text{Im}\{c_k[n]\} \sin(\Delta\omega_k[n] \cdot n) \quad (9.13)$$

with

$$\Delta\omega_k[n] = 2\pi \frac{\Delta f_k[n]}{F_s} . \quad (9.14)$$

Figure 9.12 shows the spectrograms of a short sound recording⁵ which contains strong frequency modulation, (a) without, and (b) with vibrato expansion.

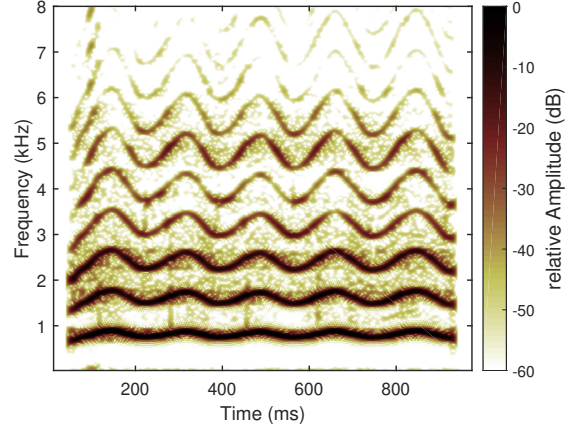
The same algorithm can also be applied for the opposite goal of vibrato suppression. By setting $\lambda = -1$, the estimated frequency difference Δf_k is then removed from the signal.

9.1.2. Temporal contrast enhancement

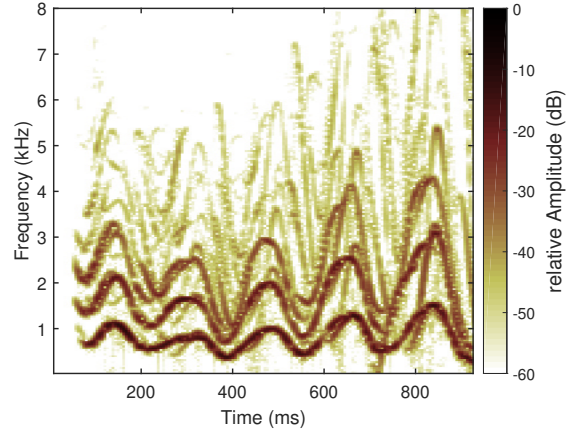
Temporal contrast enhancement is done for two reasons: (1) to make temporal structures in the sound more prominent, and (2) to compensate latency and time-smearing of the spectral contrast enhancement.

Even if sub-band processing is bypassed, the filterbank itself induces a frequency-dependent group delay which is small at high frequencies but increases towards lower frequencies (23.5 ms for the lowest band at 50 Hz).

⁵<https://github.com/pure-data/pure-data/blob/master/doc/sound/voice2.wav>



(a) original signal



(b) enhanced signal

Figure 9.12.: Spectrograms of a test signal without (a) and with (b) vibrato expansion ($\lambda=0.05$).

This frequency-dependent group delay might be acceptable for steady sounds, but it delays and smoothens any transient, transforming it to something similar to a down-chirp. Due to their broadband spectrum in combination with smoothed lateral inhibition, spectral ACE anyway effectively suppresses all transients. For maximum spectral contrast in addition with instantaneous and sharp transients, they must be detected as fast as possible from the original input signal, and mixed to the output of spectral ACE, in order to preserve them.

Transients are detected in real time by the same simple transient detection algorithm that has been used for decay prolongation (see Sec. 9.1.1.3). A 2nd-order high-pass filter with adjustable cutoff frequency f_t makes the transient detection more sensitive to high-frequency content. $s_h[n]$ is the high-

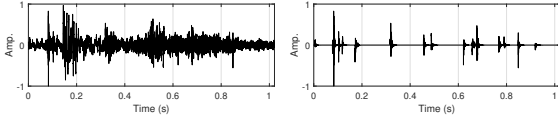


Figure 9.13.: Signal waveform without (left, Snd. S2.1) and with temporal ACE (right, Snd. S2.3).

pass-filtered version of $s[n]$. The envelope $e_t[n]$ of the transient part of the signal is estimated via the difference of a slowly decaying envelope $e_{t,d}[n]$ and a slowly rising envelope $e_{t,a}[n]$:

$$e_t[n] = \max \{ e_{t,d}[n] - e_{t,a}[n] - \nu, 0 \} \quad (9.15)$$

with threshold ν . Envelopes are computed via the two filters env_d and env_a that have been explained in Sec. 9.1.1.3 and Eq. 9.9–9.10:

$$\begin{aligned} e_{t,d}[n] &= \text{env}_d \{ s_h[n] \} \quad \text{and} \\ e_{t,a}[n] &= \text{env}_a \{ e_{t,d}[n] \}. \end{aligned} \quad (9.16)$$

The output signal of temporal ACE, $s_t[n]$, contains only the detected transients with their original amplitude:

$$s_t[n] = s[n] \cdot \frac{e_t[n]}{\text{env}_d \{ e_t[n] \}}. \quad (9.17)$$

Setting the time constants $\tau_a = 3$ ms for env_a and $\tau_d = 7$ ms for env_d , seems to work well with most of the signals we tested. Threshold ν is adjusted with respect to the overall signal level.

In sound examples Snd. 1.2–1.4, the original transients are smoothed by spectral contrast enhancement. For that reason, the original transients are extracted (Snd. S1.5) and mixed to the enhanced signals. Sound examples S1.6–1.8 are the same as Snd. S1.2–1.4, respectively, but with restored transients.

In Fig. 9.13 and Snd. S2.3, the effect of temporal contrast enhancement is demonstrated with the machine recording from Snd. S2.1. It is clearly visible that, similar to spectral sharpening, local amplitude minima are attenuated while local amplitude maxima are retained. Note that the algorithm operates on the high-pass-filtered version (cutoff frequency set to 4 kHz). The mechanic rattling thus becomes the prominent sound characteristic. A mix with the enhanced signal from spectral dynamics expansion (Snd. S2.2) leads to a spectrally and temporally enhanced signal (Snd. S2.4).

For temporal contrast enhancement, it makes no sense to apply dynamics expansion based on an absolute threshold as for spectral contrast enhancement via exponentiation—this would be a waveshaper, introducing unwanted distortion. The linear cross-fade with the dry input signal actually serves as a control for the amplitude of the residuum signal between transients.

9.1.3. Discussion

One might notice that the proposed ACE method does not explicitly include spectro-temporal contrast enhancement, e.g., temporal contrast enhancement on a sub-band level. Our hearing system does exactly that via contrast gain control in the auditory cortex, at timescales of about 100 ms (Rabinowitz et al. 2011). Rabinowitz et al. define spectro-temporal contrast as “the variation in sound pressure in each frequency band, relative to the mean”; a model can be based on the standard deviation of recent sound pressure level (Rabinowitz et al. 2011). One audible effect is that a harmonic partial which is omitted and then reintroduced may stand out perceptually for a short period of time (Summerfield et al. 1987). While this is certainly a helpful feature, it must be noted that the main objective of such adaptive gain control is to compensate the very limited dynamic range of neurons. We found that spectro-temporal contrast is anyway strong with spectral contrast enhancement alone, e.g., through a possible edge effect in case of a missing partial. Even more so, if smoothing for lateral inhibition is bypassed, together with a large ρ , a clicking transient appears whenever there is a shift of spectral energy from one band to another. Due to the group delay of the filters, however, such a transient would exhibit latency that is unacceptable for short interaction sounds.

For continuous sounds where more latency can be tolerated, it might be interesting to exaggerate amplitude modulations on a sub-band level. For that goal we tried an algorithm which expands the sub-band envelopes individually while preserving their overall envelope trend (Hoffman and Cook 2008). While originally designed to exaggerate dissonances, it is capable to enhance also low-frequency amplitude modulations. At a closer look, however, similar results could be achieved by spectral ACE alone.

Concerning spectral contrast, both methods—spectral sharpening and spectral dynamics expansion—are essential. As soon as spectral sharpening has reached its limits (i.e., what is left are local

9. Auditory contrast enhancement (ACE)

maxima only), spectral dynamics expansion can add additional contrast by suppressing all local maxima below a certain threshold.

In a parallel configuration, spectral sharpening and spectral dynamics expansion can complement each other, producing a cartoonification of the sound. This may be illustrated by the example of human speech: by lateral inhibition, speech is basically reduced to fundamental frequency and formants; consonants are attenuated. While stops/plosives could be recovered via temporal contrast enhancement, sibilants are suppressed. Exponentiation maintains or even exaggerates consonants, including sibilants; however, it has a tendency to suppress formants, so that discrimination between vowels is lost. The solution might be a combination by taking the maximum of both outputs.

Temporal contrast enhancement as implemented here works similar to a transient shaper/designer for music production. The main difference is that we try not to exaggerate transients but to attenuate everything else. A dynamics expansion would conflict with the limited dynamic range of our hearing system, and would also produce an implausible amplification of the targeted interaction sounds. The mix of spectral and temporal ACE works well for these impact sounds, but may produce quite disturbing results for more continuous stimuli such as speech.

9.1.4. Conclusions

We introduced a new method for real-time auditory contrast enhancement, targeting at interactive applications where auditory feedback is used as part of a knowledge-making process. The method is split in two parts—spectral and temporal contrast enhancement—which can be used in parallel to focus on different auditory features. Spectral ACE is achieved in two ways which both are needed for different tasks. While the first approach is based on lateral inhibition and enhances spectral sharpness, the second enhances spectral dynamics via exponentiation. In the visual domain, these would refer to edge detection and contrast, respectively. Crucial for perceptibility of the enhanced sound is decay prolongation which provides a listener with additional time for pitch impression. Transient detection was found to be sufficient for temporal contrast enhancement. The results indicate that auditory contrast can be significantly enhanced by the proposed method.

The next step is to evaluate the multitude of

parameters in order to find meaningful ranges and scalings, and ultimately reduce them to only a few intuitive controls. The next section (Sec. 9.2) describes an experiment with the goal to find a compromise for the parameters, achieving high auditory contrast while maintaining a certain degree of naturalness and plausibility of any auditory feedback.

9.2. Plausibility of enhanced auditory feedback

Like any other interface, an application of ACE in form of a hear-through system might not be accepted by the users if it lacks a plausible or natural feel (Susini et al. 2012). We obviously need to find a compromise between auditory contrast and plausibility of the augmented auditory feedback. Therefore, we designed an experiment to evaluate the perceived plausibility of the auditory feedback of physical interactions from the observer-perspective. Participants watched short video sequences and were asked to rate the plausibility of the (augmented) audio track.

9.2.1. Stimuli

The total 260 stimuli for the experiment consist of 10 videos of 3 s duration, each with 26 alternative audio tracks. Recordings are taken from the Greatest Hits dataset⁶ (Owens et al. 2016), a collection of audio/video recordings of different kinds of objects and materials being hit with a drumstick.

Selection. The selection of adequate audio/video sequences from the Greatest Hits dataset was based on several subjective and objective requirements. (1) Objects are rigid and stable (no water, leaves, gravel, etc.). (2) There is no clipping in the audio recording. (3) There should be a variability of physical properties (material, size, shape, etc.). (4) There should be a variability of sound properties (pitch, timbre, decay time, etc.). (5) There is a period of 3 seconds which contains several differently sounding hits.

We initially chose 15 stimuli from which we eliminated three that produced redundant data points in the 3D parameter space of spectral centroid, spectral bandwidth, and average decay time (see also Aramaki et al. 2009). Additional two stimuli were excluded, as they led to very saturated results in

⁶The Greatest Hits dataset:
<https://andrewowens.com/vis/>

Table 9.1.: Properties of the audio/video files used in the experiment: material and object category, original filename, starting time in the recording in s , and acoustical descriptors, i.e., spectral centroid (SCG), spectral bandwidth (SB), and average -60 dB decay time T_{60} .

ID	material/object	filename prefix	time / s	SCG / kHz	SB / kHz	T_{60} / s
0	cardboard box	2015-03-30-01-38-59	1.60	2.66	3.81	0.33
1	glass table	2015-03-30-01-29-03	9.31	7.86	5.09	0.45
2	ceramic cup	2015-02-16-16-49-06	3.24	6.89	4.25	0.28
3	forest soil	2015-09-23-16-13-51-446	21.66	3.03	4.56	0.27
4	plastic box	2015-03-30-01-54-29	2.50	2.77	3.34	0.47
5	plastic bottle	2015-03-30-02-10-02	8.81	2.69	3.50	0.34
6	wooden bar	2015-10-06-18-02-12-1	1.00	4.71	4.31	0.14
7	metal box	2015-03-30-02-18-31	0.50	4.84	4.76	0.38
8	glass bowl	2015-03-25-00-09-47	18.95	5.86	4.03	0.65
9	metal table	2015-03-30-02-22-11	19.32	7.28	4.38	0.31

a first pilot test (either always plausible or always implausible). Table 9.1 lists the remaining 10 stimuli with additional data such as material and type of the object as well as corresponding name and starting time in the dataset.

Pre-processing. The original (already synchronized) audio and video tracks were cut frame-aligned (29.97 fps) with FFmpeg⁷. Videos were re-encoded to 1080p MJPEG, in order to minimize computational effort during playback. Only the first of the two audio channels was used.

A noise sample was taken from a ‘silent’ region of each stimulus. It formed the basis of a noise profile for the built-in noise reduction of Audacity⁸ (3 bands frequency smoothing, sensitivity of 6) to reduce background noise of the stimulus by -20 dB.

Spectral ACE. Spectral dynamics are defined by ρ (sharpening) and β (expansion). We selected five pairs of values which form clearly perceivable steps from very gentle sharpening up to extreme dynamics expansion. Values are $\rho = \{2, 6, 25, 25, 25\}$, and $\beta = \{0, 0, 0, 1, 9\}$, respectively; they have been chosen in informal listening sessions. The selected decay times are equally spaced on a logarithmic scale, which is a rough approximation of their perception (Blevins et al. 2013): $T_{60} = \{0, 0.15, 0.36, 0.84, 2\}$ s. In addition to all 25 combinations of spectral dynamics and T_{60} , there was a control condition with bypassed spectral ACE ($\rho = \beta = 0$, $T_{60} = 0$ s, but still passing the filterbank). The original, unprocessed recording was not used, to prevent participants from identifying it as a reference. Remaining parameters were set to fixed values: $\sigma = 3$ ERB, $\tau_{LI} = \tau_{EX} = \tau_{DP} = 7$ ms.

Resynthesis and Transient Restoration. The noise

sample from pre-processing was again used to re-synthesize a similar noise floor in Matlab—with identical magnitude spectrum (and thus level), but with random phase spectrum. This noise is added to the output of ACE, to provide an ecologically valid scenario, circumventing the by-product of noise reduction that comes through spectral ACE. Parameters for transient restoration are set to $f_t = 4$ kHz, $\nu = -42$ dB, $\tau_d = 60$ ms, and $\tau_a = 3$ ms; the transient is added with -3 dB gain.

9.2.2. Participants, apparatus, and procedure

Eleven participants (4 female, 7 male) were recruited from university staff. Those were happy to take part without payment. They all have a professional background in sound and music computing and are thus regarded as expert listeners.

The experiment software was implemented in Pd, with GUI and video playback realized in Gem⁹. Auditory stimuli were pre-rendered in Matlab. We used a standard laptop computer with 1920×1080 pixels screen, running Debian Linux. Sound was presented via AKG K 272 HD closed-back headphones which were directly connected to the integrated Conexant CX8200 audio codec.

The 260 stimuli were presented to the participants without repetitions. Trial order was randomized between participants. Randomization was constructed of 26 consecutive rows, each including all ten videos in random order; with the restriction that the first video of a row differs from the last video of the previous row. In a next step, for each video individually, the 26 audio tracks were randomly distributed to the 26 appearances of the video.

⁷FFmpeg: <http://ffmpeg.org/>

⁸Audacity: <https://www.audacityteam.org/>

⁹Gem: <http://gem.iem.at/>

9. Auditory contrast enhancement (ACE)

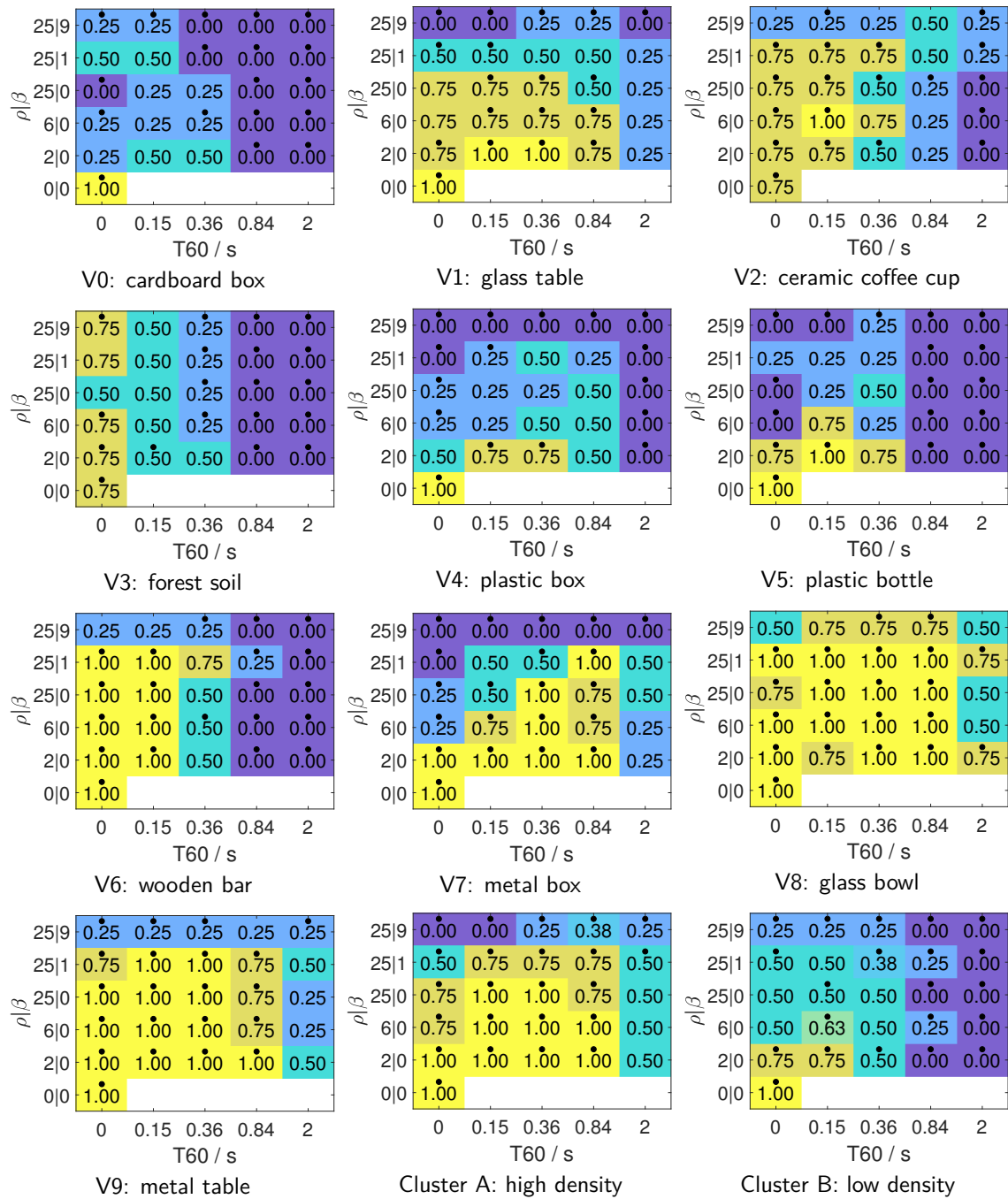


Figure 9.14.: Experiment results. Median plausibility of the 10 videos and their two main clusters, for all 26 combinations of spectral dynamics (ρ/β) and decay time T_{60} at 1 kHz, respectively. Values reach from 0 (very implausible) to 1 (very plausible). Values that are either significantly below 0.5 or significantly above or equal to 0.5 (significance level 0.05) are marked by a dot.

In each trial, the auditory-visual stimulus was automatically played back once. Participants were able to replay the stimulus in case of distraction. They were asked to rate the audio track with respect to its plausibility for the shown video. The answer had to be given on a 5-point scale, ranging from 'very implausible' (scale value 0) over 'borderline' (0.5) to 'very plausible' (1). The in-between values (0.25 and 0.75) had no label. To speed up the experiment, the answer could already be given before the end of the video. The next trial started automatically, 0.6 s after giving the answer.

Prior to the experiment, participants had to read a short summary of this procedure, including some clarifications: (1) Sound comes from an unprofessional microphone recording. (2) The task is *not* to rate authenticity or to identify the original recording. (3) The task is to quantify the plausibility of different more or less plausible alternatives. (4) Synonyms for the admittedly vague term 'plausible' are 'natural', 'convincing', and 'realistic'. One trial was randomly drawn from the whole population of trials, to serve as demonstration of the experiment interface before start. The whole experiment took about 25 min.

9.2.3. Results

From the plausibility ratings, we obtained a frequency (number of answers) for each of the five scale values and for every stimulus. We are not able to assign numbers to the plausibility ratings, as we do not know the scaling of this abstract value — we must even assume that the mental model of plausibility is different for each participant. The obtained values are therefore interpreted as purely ordinal data.

A visual inspection of the frequencies for the scale values of all 260 stimuli did not point out any multimodal distributions; we therefore assume a unimodal distribution which validates the computation of medians. These medians (between 0 and 1) are shown in Fig. 9.14 for all combinations of spectral dynamics and decay time, for the individual videos.

We consider stimuli as 'acceptable' if they reach at least 'borderline' plausibility. This is the case, when the median is at least 0.5. We tested each median with a one-tailed sign test. For medians below 0.5, we tested against the null hypothesis that it is at least to 0.5. For medians above or equal to 0.5, we tested against the null hypothesis that it is below or equal to 0.375. Those values where the null hypothesis could be rejected at a significance

level of 0.05 are highlighted in Fig. 9.14.

It is clear from this data that the independent factors spectral dynamics, decay time, and video (i.e., object/material) affect the perceived plausibility. As expected, higher values of ρ/β and T_{60} generally led to lower plausibility. The control condition was always significantly above borderline plausibility. The extreme values (top row and rightmost column) are mostly below borderline plausibility. Some videos formed very specific patterns in the 2D parameter space of ρ/β and T_{60} . The cardboard and plastic objects achieved the lowest plausibility ratings. The cardboard box was never significantly above borderline plausibility, except in the control condition. For the plastic objects, only small values of ρ/β and T_{60} were accepted. The metal box achieved highest plausibility ratings within the main diagonal, when both parameters are increased together. For ceramic, soil, and wood, only the decay prolongation seems to have a significant effect on perceived plausibility. The glass and metal objects were almost always plausible, except for extreme values.

These observations led us to do a cluster analysis, in order to validate how the individual results of the videos group together. We performed a hierarchical cluster analysis based on the pooled median values for each video. Independent of distance metrics (inner squared distance, farthest distance, shortest distance) or data (medians, means, % answers ≥ 0.5), two main clusters are emerging very clearly. Cluster A contains the four metal and glass objects, while Cluster B includes wood, soil, cardboard, plastic, and ceramics. The resulting median values of these clusters are shown in the bottom right of Fig. 9.14. For metal and glass, only the highest value of spectral dynamics and decay, respectively, is rated as implausible. The highest combination with acceptable plausibility is $\rho=25$, $\beta=1$, $T_{60}=0.84$ s (the extreme value of $T_{60}=2$ s is considered to be too large, even if it is still rated with borderline plausibility). Cluster B obviously incorporates the same tendencies as the individual videos for wood, ceramics, plastic, cardboard, and soil: low overall plausibility, weak impact of spectral contrast, and a preference for the diagonal. Highest acceptable values are $\rho=25$, $\beta=0$, $T_{60}=0.15$ s.

9.2.4. Discussion

Some of the above results might allow us to draw conclusions on the participant's mental model of plausibility in the context of the experiment. In general, we consider sensory feedback as plausible

9. Auditory contrast enhancement (ACE)

if it is “conceptually consistent with what is known to have occurred in the past” (Connell and Keane 2006).

The cluster analysis clearly divides the stimuli in high-density materials (glass and metal) and low-density materials (plastic, wood, ceramics, soil, and cardboard). Such grouping into gross density categories has already been observed in other listening experiments from the literature; people can hardly discriminate materials within those groups (see Giordano and McAdams 2006 and Sec. 2.2.1).

Within the high-density materials, the plausibility of the metal table and glass bowl was almost unaffected by ACE, as these exhibit anyway strong spectral sharpness and long decay. Plausibility suffered only if disturbing artifacts occurred with extreme values of spectral dynamics (jumps between partials) and long decay time (sharp partials transform into band-pass noise). Interestingly, the glass bowl was partly rated above borderline with the 2 s decay prolongation, although it is clearly visible that it is placed on its highly damping plastic cover. This fact was easily ignored by the participants. The metal box is actually a paper dispenser which is included in two versions (filled and empty) in the Greatest Hits dataset. We selected the empty one; however, its inner structure and content is not completely visible, which might be an explanation for its insensitivity to decay prolongation. In a physically plausible scenario, this would also come together with increased spectral dynamics.

Within the low-density materials, the plastic bottle and cardboard box achieved exceptionally low plausibility ratings. A possible explanation might be the low frequency of the strongest resonance. For the cardboard box, it lies at 72 Hz; the group delay of the corresponding gammatone filter ($f_c = 71$ Hz) is already 21.9 ms — unacceptable for plausible auditory feedback. With increased spectral contrast, this delay gets more salient. For the forest soil, it is obvious that the participants could not find a physical explanation for a long decay time above 0.1 s. However, the soil was insensitive to spectral ACE, as even the original recording exhibits distinct pitch at different positions.

An objective measure of spectral contrast can be obtained via the entropy-based measure of spectral flatness SF (Madhu 2009). Spectral contrast is then $SC = 1 - SF$. Overall spectral contrast \overline{SC} is derived from N 50 %-overlapping Hann-windowed signal blocks of 1024 samples. Spectral contrast SC_n of each block n is weighted with its energy



Figure 9.15.: Schematic drawing of the proposed ACE hear-through system.

E_n , respectively:

$$\overline{SC} = \frac{1}{\sqrt{N}} \frac{\sum_n \sqrt{E_n} SC_n}{\sqrt{\sum_n E_n}}. \quad (9.18)$$

Compared to the control condition with bypassed ACE, measured spectral contrast increases monotonically for increasing ρ and β (averaged over video and decay; from 51 % to 87 %) as well as for increasing T_{60} (averaged over video and spectral dynamics; from 38 % to 107 %). While maintaining at least borderline median plausibility (significantly ≥ 0.5), the average spectral contrast is enhanced by 49 % for cluster A and 26 % for cluster B. This confirms that auditory contrast can be plausibly enhanced by the proposed ACE method.

9.3. The ACE hear-through system

The technical setup of the ACE hear-through system is based on earphones with integrated binaural microphones, as illustrated in Fig. 9.15. We use the Roland CS-10EM microphone/earphone combination. In principle, the software runs on any standard laptop; however, due to the demands on latency, a workstation PC with RME HDSPe MADI FX audio interface is used for the first prototype. ACE is implemented in SuperCollider under Debian Linux. Measured round-trip latency of this setup, from the loudspeaker of one earpiece to its integrated microphone, is 3.6 ms (at 48 kHz sampling rate). This seems to be sufficient; less than 10 ms is usually

recommended for auditory-tactile environments as well as for hearing aids (Altinsoy 2012; Bramsløw 2010). Open canal hearing aids, however, require even lower latency below 5 ms (Herbig and Chalupper 2010).

The system does not alter inter-aural time differences, but inter-aural level differences might be altered quite unpredictable, if both channels are processed individually. As localization is anyway bad with such hear-through systems (Marentakis and Liepins 2014), we decided to sum both microphone signals to a monophonic signal before being processed through the algorithm for real-time intra-stimulus ACE. The final envelopes, however, are then applied to the original binaural signal for left and right ear individually for preserving the original inter-aural time and level differences.

We build upon the gammatone filter design by Hohmann (Hohmann 2002); its SuperCollider implementation¹⁰ has been adapted to output real and imaginary part of the signal. The ACE hear-through system provides a reduced GUI which should be sufficient for most tasks. However, expert users are always able to switch to an extended GUI which allows tuning of all the described parameters of spectral and temporal ACE.

Demonstration videos with different materials are available in Source 9.2.



9.4. Conclusions and outlook

We presented a hear-through system for intra-stimulus auditory contrast enhancement. It is intended to be used wherever auditory feedback is part of a knowledge-making process. Two main parameters control spectral sharpening and spectral dynamics expansion, and decay prolongation. Both parameters have a positive effect on measured spectral contrast. Perceptual plausibility of the resulting sounds was measured in an experiment with auditory-visual stimuli. For all except one of the 10 tested sounding objects, measured auditory contrast could be increased while maintaining the amount of plausibility that is necessary for a natural user interface. Besides the rather technical applications mentioned in the introduction, we see great potential also in sound cartoonification and sound transformations for artistic or musical purpose.

Future work might include a user study, where the ACE hear-through system is evaluated in a real-time

¹⁰gammatone UGen in *AuditoryModeling* UGens in *SC3 Plugins*: <https://github.com/supercollider/sc3-plugins>

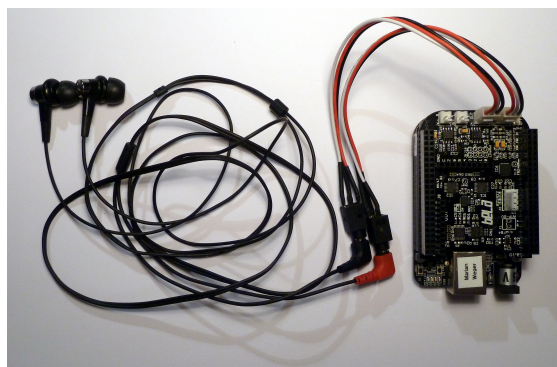


Figure 9.16.: Mobile ACE based on Bela.

scenario that simulates its primary target application: percussion. It is further planned to optimize the implementation so that it runs on mobile devices such as the Bela platform (see Fig. 9.16), or maybe even as a smartphone app. A fusion with the powerful non-real-time methods for inter-stimulus ACE (Hermann and Weger 2019) into a single mobile artifact would create a versatile tool for interactive sound exploration.

The source code of the ACE hear-through system is published as free software (Source 9.3).

</>

Bibliography

- Altinsoy, M Ercan (2012). “The Quality of Auditory-Tactile Virtual Environments”. In: *Journal of the Audio Engineering Society* 60.1, pp. 38–64.
- Aramaki, Mitsuko et al. (2009). “Timbre Perception of Sounds from Impacted Materials: Behavioral, Electrophysiological and Acoustic Approaches”. In: *Computer Music Modeling and Retrieval (CMMR)*. Vol. 5493. Springer, pp. 1–17. DOI: 10.1007/978-3-642-02518-1_1.
- Baer, Thomas, Brian C. J. Moore, and Stuart Gatehouse (1993). “Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times”. In: *Journal of Rehabilitation Research* 30.1, pp. 49–72.
- Békésy, G. von (1962). “Lateral inhibition of heat sensations on the skin”. In: *Journal of applied physiology* 17.6, pp. 1003–1008.
- Bishop, P. James (Mar. 1961). “A bibliography of Auenbrugger’s ‘Inventum Novum’ (1761)”. In: *Tubercle* 42.1, pp. 78–90. DOI: 10.1016/S0041-3879(61)80023-8.

9. Auditory contrast enhancement (ACE)

- Blevins, Matthew G et al. (2013). "Quantifying the just noticeable difference of reverberation time with band-limited noise centered around 1000 Hz using a transformed up-down adaptive method". In: *International Symposium on Room Acoustics (ISRA)*. Toronto, Canada.
- Boers, P.M. (1980). "Formant enhancement of speech for listeners with sensorineural hearing loss." In: *IPO Annual Progress Report 15*, pp. 21–28.
- Bramsløw, Lars (Sept. 2010). "Preferred signal path delay and high-pass cut-off in open fittings". In: *International Journal of Audiology* 49.9, pp. 634–644. DOI: 10.3109/14992021003753482.
- Connell, Louise and Mark T. Keane (Jan. 2, 2006). "A Model of Plausibility". In: *Cognitive Science* 30.1, pp. 95–120. DOI: 10.1207/s15516709cog0000_53.
- Coren, Stanley et al. (Dec. 1988). "A method to assess the relative contribution of lateral inhibition to the magnitude of visual-geometric illusions". In: *Perception & Psychophysics* 43.6, pp. 551–558. DOI: 10.3758/BF03207743.
- Ertel, Paul Y., Merle Lawrence, and Wonjin Song (1971). "Stethoscope acoustics and the engineer: Concepts and problems." In: *Journal of the Audio Engineering Society* 19.3, pp. 182–186.
- Giordano, Bruno L. and Stephen McAdams (2006). "Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates". In: *Journal of the Acoustical Society of America (JASA)* 119.2, pp. 1171–1181. DOI: 10.1121/1.2149839.
- Herbig, R. and R. Chalupper (2010). "Acceptable processing delay in digital hearing aids." In: *Hearing Review* 17.1, pp. 28–31.
- Hermann, Thomas, Andy Hunt, and John G Neuhoff (2011). *The sonification handbook*. Berlin: Logos Verlag. ISBN: 978-3-8325-2819-5.
- Hermann, Thomas and Marian Weger (June 2019). "Data-driven Auditory Contrast Enhancement for Everyday Sounds and Sonifications". In: *International Conference on Auditory Display (ICAD)*. Newcastle upon Tyne, UK, pp. 83–90. DOI: 10.21785/icad2019.005.
- Hoffman, Matthew and Perry Cook (2008). "Real-Time Dissonancizers: Two Dissonance-Augmenting Audio Effects". In: *International Conference on Digital Audio Effects (DAFx)*. Espoo, Finland.
- Hohmann, V (2002). "Frequency analysis and synthesis using a Gammatone filterbank". In: *Acta Acustica united with Acustica* 88, pp. 433–442.
- Houtgast, T. (June 1972). "Psychophysical Evidence for Lateral Inhibition in Hearing". In: *The Journal of the Acoustical Society of America* 51.6, pp. 1885–1894. DOI: 10.1121/1.1913048.
- Hutchins, Carleen Maley (1962). "The Physics of Violins". In: *Scientific American* 207.5, pp. 78–93.
- Koumura, Takuya and Shigeto Furukawa (Dec. 2017). "Context-Dependent Effect of Reverberation on Material Perception from Impact Sound". In: *Scientific Reports* 7.1. DOI: 10.1038/s41598-017-16651-4.
- Kral, A. and V. Majernik (1996). "On lateral inhibition in the auditory system.pdf." In: *Gen. Physiol. Biophys.* 15, pp. 109–127.
- Madhu, N. (2009). "Note on measures for spectral flatness". In: *Electronics Letters* 45.23. DOI: 10.1049/el.2009.1977.
- Marentakis, Georgios and Rudolfs Liepins (2014). "Evaluation of hear-through sound localization". In: *Conference on Human factors in computing systems (CHI)*. Toronto, Canada: ACM, pp. 267–270. DOI: 10.1145/2556288.2557168.
- Moore, Brian C. J. and Brian R. Glasberg (1996). "A Revision of Zwicker's Loudness Model". In: *Acta Acustica united with Acustica* 82, pp. 335–345.
- Noisternig, Markus (2017). „Breitbandige Signalaufbereitung in Ein- und Mehrkanal-Mikrofonanwendungen“. Diss. Universität für Musik und darstellende Kunst Graz.
- Nuland, Sherwin B. (2005). *Doctors: The History of Scientific Medicine Revealed Through Biography*. Chantilly, VA: The Great Courses.
- Owens, Andrew et al. (June 2016). "Visually Indicated Sounds". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE, pp. 2405–2413. DOI: 10.1109/CVPR.2016.264.
- Pantev, C. et al. (Apr. 2004). "Lateral inhibition and habituation of the human auditory cortex". In: *European Journal of Neuroscience* 19.8, pp. 2337–2344. DOI: 10.1111/j.0953-816X.2004.03296.x.
- Patterson, Roy et al. (1987). "An efficient auditory filterbank based on the gammatone function". In: *Speech-Group meeting of the Institute of Acoustics on Auditory Modelling*. RSRE, Malvern.
- Rabinowitz, Neil C. et al. (June 2011). "Contrast Gain Control in Auditory Cortex". In: *Neuron* 70.6, pp. 1178–1191. DOI: 10.1016/j.neuron.2011.04.030.

- Stone, Michael A. and Brian C. J. Moore (1992). "Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligibility and quality". In: *The Journal of Rehabilitation Research and Development* 29.2. DOI: 10.1682/JRRD.1992.04.0039.
- Summerfield, Quentin, Andrew Sidwell, and Tony Nelson (Mar. 1987). "Auditory enhancement of changes in spectral amplitude". In: *The Journal of the Acoustical Society of America* 81.3, pp. 700–708. DOI: 10.1121/1.394838.
- Susini, Patrick et al. (May 2012). "Naturalness influences the perceived usability and pleasantness of an interface's sonic feedback". In: *Journal on Multimodal User Interfaces* 5.3, pp. 175–186. DOI: 10.1007/s12193-011-0086-0.
- Weger, Marian, Thomas Hermann, and Robert Höldrich (June 2019). "Real-time Auditory Contrast Enhancement". In: *International Conference on Auditory Display (ICAD)*. Newcastle upon Tyne, UK, pp. 254–261. DOI: 10.21785/icad2019.026.
- Weger, Marian and Robert Höldrich (2019). "A hear-through system for plausible auditory contrast enhancement". In: *Proceedings of Audio Mostly*. Nottingham, UK: ACM, pp. 1–8. DOI: 10.1145/3356590.3356593.
- Yang, Jun, Fa-Long Luo, and Arye Nehorai (2003). "Spectral contrast enhancement: Algorithms and comparisons". In: *Speech Communication* 39, pp. 33–46.

10. General conclusions and outlook

The primary goal of this dissertation was to find ways for projecting digital information into our physical world by the use of sound — under the condition that no further sound events are added to the everyday acoustic environment. This rather strict premise stays in conflict with traditional methods of sonification and obliges us to modulate already existing sounds for conveying additional information — this is what we call auditory augmentation (in the strict sense). In the more loose sense, not only physical objects but also user actions may be augmented by sound. In either way, we argued that the auditory augmentation needs to stay within the limits of plausibility for the given physical interaction, in order to get accepted by possible users. Furthermore, we decided that the augmentation must not deteriorate the usability of the original physical object. The latter condition was rather easy to fulfill: if the augmented auditory feedback (1) still conveys the usability-relevant information and (2) does not disturb other activities (e.g., by masking relevant auditory feedback), then we assume it to be usable. For plausibility, it is more complicated: we still do not know the limits of plausibility. Most of the auditory augmentations that have been presented in this thesis (except some prototypes in Sec. 7.3) successfully circumnavigated this problem by choosing physically plausible parameters in the first place (e.g., material, size, aspect ratio, room reverberation time, etc.), and by modulating these within ecologically valid ranges.

Based on the results of an interdisciplinary workshop on auditory augmentation (SBE4, Ch. 7), we were able to draw conclusions on the prospects and limits of auditory augmentations in general. Auditory augmentations are structured around three main components: the physical object, the task of the user, and the sound (i.e., auditory feedback). The type of auditory augmentation is defined by the connection between these components. Task-centered auditory augmentations are based around a main task which generates data that is displayed by sound. In sound-centered auditory augmentations, the task generates auditory feedback that is modulated by the data. Data-centered auditory augmentations start with data that are displayed by

sound which is finally modulated by the task. In summary, auditory augmentation requires a primary task (not data exploration) that is not disturbed by the augmentation. In return, it adds a secondary monitoring or data exploration task. Furthermore, auditory augmentation favors a tight connection between an action and the resulting sound as well as a metaphor which links the sound to the physical state of the object. Finally, auditory augmentation works best if it is coupled to data which depend on time; likewise it cannot be consumed passively at a glance, but requires active participation by the user for a certain duration.

The main sonification platform described within this thesis is AltAR/table (Ch. 4): an auditory augmentation of a rectangular plate that simulates the sonic behavior of any other rectangular and isotropic or orthotropic plate by setting physical parameters. Users can map different physical models to different regions of the interface plate, while the borders between regions as well as the model parameters may be modulated in real time, e.g., by external data, for providing an auditory display of digital information. The display is calm and unobtrusive (no additional sound events are added), conveys its underlying information ambiently in the background, e.g., while writing or placing objects, and is always at hand if required: by knocking on the table. For hiding the technical apparatus from the user, sensors and actuators are mounted below the interface plate. For using the same plate as contact microphones and structure-borne exciters, several methods of feedback prevention, subtraction, and suppression had to be applied. While position tracking is not considered as an essential feature for a final product, hand damping, e.g., through weight sensors, was found to be an indispensable feature for providing a natural user experience.

The second sonification platform is the augmented room. In Ch. 7 we concluded that a room can indeed be seen as an object for plausible auditory augmentation. While its reverberation affects all sounds that reach it, we are quite good at separating it from the original sound event (Koumura and Furukawa 2017). In Sec. 8.2 and 8.3, we showed that variable room acoustics can serve as a plausible,

10. General conclusions and outlook

calm, and salient communication channel for data monitoring, even while pursuing everyday activities such as preparing coffee or listening to the radio. Technically, all sound within the room is captured by microphones, filtered by artificial reverberation, and played back by loudspeakers.

Both platforms (room and table) draw on a physical sound model to synthesize natural and physically plausible auditory feedback. While readily available tools have been applied for virtual room acoustics (Sec. 8.3.2.2), a custom physical model was implemented for the rectangular plate of the table (Ch. 3). Instead of simulating the physical system directly, e.g., via digital waveguides, a modal synthesis approach was used. The therefore computed intermediate sound parameters which characterize the modes were additionally used for informed feedback suppression. The model covers any thin rectangular plate of orthotropic or isotropic material, and with any combination of three boundary conditions (free, hinged, clamped). It includes different forms of damping (viscous, viscoelastic, and thermoelastic), indentation hardness of the plate surface, and frequency-dependent radiation efficiency. The physical model not only serves as sound synthesis engine, but also allows a deeper understanding of how physical information is encoded into sound, i.e., how physical parameters map to sound parameters.

In the opposite direction, the physical model allows us to draw conclusions on how physical parameters can be decoded from auditory feedback. In theory, the inverse of the physical model perfectly describes this relationship. In practice, however, the inverse model represents an underdetermined system: the same sound may be achieved by different combinations of physical parameters. We developed a novel algorithm for automated identification of length, width, and material of rectangular plates, based on impact sounds (Sec. 5.1). Length, width, and material category of unknown plates could be estimated with high precision, provided that the damping is sufficiently low. Some parameters such as thickness, elasticity, and density, however, were ambiguous due to underdetermination. We therefore proposed a Bayesian interpretation of the results. It incorporates additional information such as context or physical limitations for assigning probabilities to all combinations of the ambiguous physical parameters. This algorithm for model-based robotic perception might be useful in practical applications such as non-destructive inspection or handling of hazardous waste. In addition, it helps us to understand how physical information is extracted from

sound by human listeners.

The perception of physical parameters of rectangular plates by human listeners was evaluated in two experiments.

In a multisensory experiment (Sec. 5.2), participants explored the sound of unknown rectangular plates via percussion on a simulator based on the AltAR/table platform. They were asked to estimate lengths and aspect ratios in reliance on auditory feedback. From a sonification designer's perspective, it is convenient to demand a certain probability of superiority P_s , concerning the discrimination between parameter levels.

At $P_s = 90\%$, the participants achieved 2 discriminable levels of aspect ratio (independent of the plate's material), and between 2 and 3 discriminable levels of length (for plates of glass and wood, respectively). When both parameter dimensions were unknown (i.e., in a 2D auditory display), the identification performance was slightly inferior due to confusion between plates of equal surface area (e.g., small and compact vs. larger and longish).

In a listening experiment (Sec. 5.3), participants had to identify material and aspect ratio of rectangular plates. The presented impact sounds were synthesized on the basis of three meta-parameters: rigidity, metallicity, and aspect ratio. In absolute judgments (comparable to a 3D auditory display), participants were able to discriminate between 2 levels of rigidity and metallicity, respectively, at $P_s = 75\%$, but could hardly discriminate between aspect ratios.

Both augmented table and augmented room expect absolute magnitude estimations of sound parameters such as frequencies, amplitudes, decay times, etc. Except some individuals with absolute pitch perception, humans are generally bad at this task. Notwithstanding our exceptional capabilities concerning relative judgments, absolute judgments are limited to a maximum of about 7 levels—independent of the specific sensory channel (Miller 1956). Perception-wise, the strongest limiting factor for auditory displays that demand magnitude estimation is thus the capacity of our short-term memory. With the augmented room (RadioReverb experiment in Sec. 8.3) we apparently reached this limit, even in a background task scenario. Fortunately, the limit can be exceeded by adding further dimensions (Pollack and Ficks 1954). Unfortunately, this research line is relatively unknown within the sonification community; and

even within this thesis we only discovered it after carrying out our experiments. AltAR/table therefore did not even reach the “magical number seven” due to one major problem: the individual parameter dimensions were (sonically) not entirely independent. Participants therefore could not evaluate the respective sound parameters independently, but were forced to memorize individual parameter combinations within their short-term memory. Now that we know the possible solution, we would try to design auditory augmentations with higher dimensionality and higher orthogonality between dimensions; at the cost of perceptual resolution within dimensions, if necessary.

While we focused more on magnitude estimation when evaluating auditory augmentations, we did not explicitly consider the two main strengths of our auditory system: relative judgments and high temporal resolution. This focus comes from the premise of plausible auditory augmentation: it may be acceptable if a physical object morphs slowly (within minutes) into another material, but it seems entirely implausible if the material oscillates (within milliseconds) between materials. The latter case was rejected beforehand, as it would shift our focus from ecologically valid physical parameters to arbitrary sound parameters. While perceptual parameters such as roughness would be more salient, it might be difficult to integrate them into physically plausible auditory feedback. A central idea was that we are already accustomed to natural interaction sounds: we assumed that it would require less training for understanding physically plausible auditory displays than implausible ones that are tuned to maximum saliency. The latter, however, are regarded as the first choice in a professional environment such as medical applications where plausibility is only secondary. Another assumption was that we are able to absolutely identify specific physical properties such as material, based on our knowledge from everyday life. This assumption was partly confirmed by the relatively good discrimination performance between non-metals and metals (Sec. 5.3.4.2). Nevertheless, we could observe the same difficulties that have been reported by the literature (Sec. 2.2.1): near perfect discrimination between gross density categories (e.g., wood vs. metal), but poor discrimination within these categories (e.g., glass vs. metal).

A secondary goal of this dissertation was to find ways for enhancing the information that is already encoded in everyday auditory feedback. This was achieved by several types of auditory contrast en-

hancement (ACE, Ch. 9)—especially spectral contrast enhancement (Sec. 9.1.1). Spectral contrast enhancement transfers the concepts of edge detection and visual contrast from image processing to spectral manipulations of sound. While these two approaches are anyway performed already by our auditory system (via lateral inhibition and contrast gain control, respectively), the algorithm allows more extreme settings, serving as “auditory magnifying glass”. In order to give listeners more time to perceive short impact sounds, ACE allows to prolongate the exponential decay of natural impact sounds while drawing only on their inherent signal content. In general, we differentiate between intra-stimulus ACE (i.e., making intrinsic features of a sound more prominent) and inter-stimulus ACE (i.e., making differences between two sets of sounds more prominent). Under certain circumstances, e.g., when unique features of individual sounds are masked by other sound components, the presented algorithm for intra-stimulus ACE likewise enhances inter-stimulus contrast.

ACE has been evaluated based on auditory-visual recordings of everyday objects being struck by a drumstick (Sec. 9.2). For materials of high density, we achieved an increase of measured spectral contrast by 49% while staying within the range of acceptable plausibility between audition and vision. For low-density materials, the gain in measured spectral contrast was still 26%. In some situations, ACE might even be useful as a general purpose effect at the end of the signal chain for tuning or mastering arbitrary sonifications. Likewise, besides shaping the sound for enhanced perceptibility, ACE might be directed even to more artistic applications for achieving cartoonifications of sounds through the parallel combination of spectral sharpening (edge detection) and spectral dynamics expansion (contrast).

The real-time implementation of ACE in form of a hear-through system (similar to a closed-canal hearing aid) seems promising, although its experimental evaluation is still pending (Sec. 9.3). The working prototype is based on earphones with integrated binaural microphones and targets mobile applications where exploratory listening is performed. However, it still requires a heavy computer for achieving an acceptable round-trip latency below 10 ms.¹

With CardioScope (Sec. 8.4), we showed that even well-established tools for medical diagnosis can

¹A computationally more efficient port to microcomputers such as BeagleBone (via Bela or Raspberry Pi, implemented in C/C++, is currently in the making.

benefit from auditory contrast enhancement. CardioScope is an augmented stethoscope that makes specific temporal regions of interest within the cardiac cycle more salient, with the help of data acquired from the electrocardiogram (ECG). It can be regarded as a special case of real-time ACE, where the sound is shaped with respect to a second information channel that is processed in parallel.

The plausibility of auditory feedback constituted a central research topic of this thesis. It is a pity that we still lack a proper theory that might predict the plausibility of sound with respect to a given physical action and context. Nevertheless, we already developed a tool for exploring it. Schrödinger's box (Ch. 6) is a shiny black cube with one single affordance: striking it with a small mallet. The resulting augmented auditory feedback is created by battery-powered electronics that are hidden inside. Novel methods for real-time onset detection and sound playback allow the exceptionally low round-trip latency far below 5 ms. The original and synthetic auditory feedback thus fuse to a single sound object for achieving plausible auditory feedback. The sound synthesis draws on recorded samples and covers a large timbre space reaching from perfectly plausible to obviously implausible auditory feedback. In future experiments, it is planned to explore the limits of plausibility for the given task and object (striking an unknown black box), and to derive a theoretical model for predicting the plausibility of auditory feedback, based on closely related concepts such as the plausibility analysis model (PAM) by Connell and Keane (2006) or the causal certainty measure by Ballas and Sliwinski (1986). In addition to the original intention, Schrödinger's box turns out to be a great platform for teaching and prototyping interactive sonifications or sonic interactions.²

None of the questions or problems that were raised within this thesis have been completely answered or ultimately solved. It seems that an even larger number of newly unsolved questions has been added. The research field of auditory augmentation proved to be way larger than what could be covered within the scope of this thesis. This underlines the relevancy of the topic, not only concerning this work, but also further research to come. While none of the presented sonifications describes a final product, the gathered experience and even the mistakes that

have been made may form the basis of more sophisticated auditory augmentations or even practical applications in the future. For further reading, the articles that formed the foundation this dissertation or emerged from it are listed in the end on p. 223.

Bibliography

- Ballas, J. A. and M. J. Sliwinski (Nov. 1, 1986). *Causal Uncertainty in the Identification of Environmental Sounds*. ONR-86-1. Fort Belvoir, VA: Defense Technical Information Center. DOI: 10.21236/ADA175228.
- Connell, Louise and Mark T. Keane (Jan. 2, 2006). "A Model of Plausibility". In: *Cognitive Science* 30.1, pp. 95–120. DOI: 10.1207/s15516709cog0000_53.
- Koumura, Takuya and Shigeto Furukawa (Dec. 2017). "Context-Dependent Effect of Reverberation on Material Perception from Impact Sound". In: *Scientific Reports* 7.1. DOI: 10.1038/s41598-017-16651-4.
- Miller, George A. (Mar. 1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological Review* 63.2, pp. 81–97. DOI: 10.1037/h0043158.
- Pollack, Irwin and Lawrence Ficks (Mar. 1954). "Information of Elementary Multidimensional Auditory Displays". In: *The Journal of the Acoustical Society of America* 26.2, pp. 155–158. DOI: 10.1121/1.1907300.

²If falling back to two channels instead of four, and loudspeakers instead of exciters, it can be constructed with less effort and from less expensive components. For integration into smaller physical objects, a port to microcontrollers, e.g., via CircuitPython, is considered.

A. Appendix to the physical model

The computations of approximate characteristic beam functions and frequency factors of rectangular thin plates are based on the equations provided by Warburton (1954).

A.1. Approximate characteristic beam functions

hinged [hh] If hinged on both ends, the characteristic beam function is

$$\theta_m^{hh} = \sin \left((m-1) \pi \frac{x}{l_x} \right) , \quad (\text{A.1})$$

for $m = \{2, 3, 4, \dots\}$, where m is the number of nodal lines, and x/l_x is the normalized spatial position on the beam. In m , also possible nodes on hinged or clamped edges of the plate are counted, as well as the one in the center for the rotation-only mode that occurs with free edges.

clamped/clamped [cc] If clamped on both ends, the characteristic beam function is

$$\theta_m^{cc}(x) = \begin{cases} \cos \left(\gamma_1 \left[\frac{m}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) + k_1^+ \left[\frac{m}{2} \right] \cosh \left(\gamma_1 \left[\frac{m}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) , & \text{for } m = \{2, 4, 6, \dots\} \\ \sin \left(\gamma_2 \left[\frac{m+1}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) + k_2^- \left[\frac{m+1}{2} \right] \sinh \left(\gamma_2 \left[\frac{m+1}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) , & \text{for } m = \{3, 5, 7, \dots\} \end{cases} , \quad (\text{A.2})$$

with

$$k_1^+ [n] = \frac{\sin \left(\frac{\gamma_1 [n]}{2} \right)}{\sinh \left(\frac{\gamma_1 [n]}{2} \right)} , \quad (\text{A.3})$$

$$k_2^- [n] = - \frac{\sin \left(\frac{\gamma_2 [n]}{2} \right)}{\sinh \left(\frac{\gamma_2 [n]}{2} \right)} , \quad (\text{A.4})$$

and γ_1 and γ_2 being the zeros of the equations

$$\tan \left(\frac{1}{2} \gamma_1 \right) + \tanh \left(\frac{1}{2} \gamma_1 \right) = 0 \quad \text{and} \quad (\text{A.5})$$

$$\tan \left(\frac{1}{2} \gamma_2 \right) - \tanh \left(\frac{1}{2} \gamma_2 \right) = 0 . \quad (\text{A.6})$$

free/free [ff] For free boundary on both ends, the characteristic beam functions is

$$\theta_m^{ff}(x) = \begin{cases} 1 & , \text{ for } m = 0 \\ 1 - 2 \frac{x}{l_x} & , \text{ for } m = 1 \\ \cos \left(\gamma_1 \left[\frac{m}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) + k_1^- \left[\frac{m}{2} \right] \cosh \left(\gamma_1 \left[\frac{m}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) & , \text{ for } m = \{2, 4, 6, \dots\} \\ \sin \left(\gamma_2 \left[\frac{m+1}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) + k_2^+ \left[\frac{m+1}{2} \right] \sinh \left(\gamma_2 \left[\frac{m+1}{2} \right] \left(\frac{x}{l_x} - \frac{1}{2} \right) \right) & , \text{ for } m = \{3, 5, 7, \dots\} \end{cases} , \quad (\text{A.7})$$

A. Appendix to the physical model

with

$$k_1^- [n] = -\frac{\sin\left(\frac{\gamma_1[n]}{2}\right)}{\sinh\left(\frac{\gamma_1[n]}{2}\right)}, \quad (\text{A.8})$$

$$k_2^+ [n] = \frac{\sin\left(\frac{\gamma_2[n]}{2}\right)}{\sinh\left(\frac{\gamma_2[n]}{2}\right)}, \quad (\text{A.9})$$

and γ_1 and γ_2 being the zeros of Eq. A.5-A.6.

clamped/free [cf] and [fc] For a cantilever beam that is clamped on one end and free on the other end, the characteristic beam function is

$$\theta_m^{\text{cf}}(x) = \cos\left(\gamma_3[n]\frac{x}{l_x}\right) - \cosh\left(\gamma_3[n]\frac{x}{l_x}\right) - k_3[n] \left(\sin\left(\gamma_3[n]\frac{x}{l_x}\right) - \sinh\left(\gamma_3[n]\frac{x}{l_x}\right) \right), \quad (\text{A.10})$$

for $m = \{1, 2, 3, \dots\}$, with

$$k_3[n] = \frac{\sin(\gamma_3[n]) - \sinh(\gamma_3[n])}{\cos(\gamma_3[n]) - \cosh(\gamma_3[n])}, \quad (\text{A.11})$$

and γ_3 being the zeros of the equation

$$\cos(\gamma_3) \cosh(\gamma_3) = -1. \quad (\text{A.12})$$

The reverse beam [fc] is then

$$\theta_m^{\text{fc}}(x) = \theta_m^{\text{cf}}\left(1 - \frac{x}{l_x}\right). \quad (\text{A.13})$$

clamped/hinged [ch] and [hc] If clamped on one end and hinged on the other end, the characteristic beam function is

$$\theta_m^{\text{ch}}(x) = \sin\left(\gamma_2[m-1]\frac{x}{2l_x}\right) + k_2^- [m-1] \sinh\left(\gamma_2[m-1]\frac{x}{2l_x}\right), \quad \text{for } m = \{2, 3, 4, \dots\} \quad (\text{A.14})$$

with k_2^- from Eq. A.4 and γ_2 being the zeros of Eq. A.6. The reverse beam [hc] is then

$$\theta_m^{\text{hc}}(x) = \theta_m^{\text{ch}}\left(1 - \frac{x}{l_x}\right). \quad (\text{A.15})$$

free/hinged [fh] and [hf] If free on one end and hinged on the other end, the characteristic beam function is

$$\theta_m^{\text{fh}}(x) = \begin{cases} 1 - \frac{x}{l_x}, & \text{for } m = 1 \\ \sin\left(\gamma_1[m-1]\left(\frac{x}{2l_x} - \frac{1}{2}\right)\right) + k_1^- [m-1] \sinh\left(\gamma_1[m-1]\left(\frac{x}{2l_x} - \frac{1}{2}\right)\right), & \text{for } m = \{2, 3, 4, \dots\} \end{cases}, \quad (\text{A.16})$$

with k_1^- from Eq. A.8 and γ_1 being the zeros of Eq. A.5. The reverse beam [hf] is then

$$\theta_m^{\text{hf}}(x) = \theta_m^{\text{fh}}\left(1 - \frac{x}{l_x}\right). \quad (\text{A.17})$$

Pre-computations Equations A.5, A.6, and A.12 are solved numerically in Matlab. A pre-computation of 100 zeros of each equation is enough to render at least 100 modes in 1D (bar) or 10 000 modes in 2D (plate).

The mode shapes are pre-computed for normalized length $l_x=1$ m and width $l_y=1$ m, a default aspect ratio of $r_a=2$, and a spatial resolution of 30 points per nodal line. In case of in total 512 modes ($M=32$ modes in x -direction and $N=16$ modes in y -direction), this results in a spatial grid of 960×480 points.

A.2. Approximate frequency factors of rectangular thin plates

hinged/hinged [hh]

$$G = n, \quad H = J = n^2, \quad \text{for } n \geq 2 \quad (\text{A.18})$$

clamped/clamped [cc]

$$G = \begin{cases} 1.506, & \text{for } n = 2 \\ n - 0.5, & \text{for } n > 2 \end{cases} \quad (\text{A.19})$$

$$H = J = \begin{cases} 1.248, & \text{for } n = 2 \\ (n - 0.5)^2 \cdot \left(1 - \frac{2}{(n-0.5)\pi}\right), & \text{for } n \geq 3 \end{cases} \quad (\text{A.20})$$

free/free [ff]

$$G = \begin{cases} 0, & \text{for } n = \{0, 1\} \\ 1.506, & \text{for } n = 2 \\ n - 0.5, & \text{for } n > 2 \end{cases} \quad (\text{A.21})$$

$$H = \begin{cases} 0, & \text{for } n = \{0, 1\} \\ 1.248, & \text{for } n = 2 \\ (n - 0.5)^2 \cdot \left(1 - \frac{2}{(n-0.5)\pi}\right), & \text{for } n \geq 3 \end{cases} \quad (\text{A.22})$$

$$J = \begin{cases} 0, & \text{for } n = \{0, 1\} \\ 1.248, & \text{for } n = 2 \\ (n - 0.5)^2 \cdot \left(1 - \frac{6}{(n-0.5)\pi}\right), & \text{for } n \geq 3 \end{cases} \quad (\text{A.23})$$

clamped/free [cf] or [fc]

$$G = \begin{cases} 0.579, & \text{for } n = 1 \\ 1.494, & \text{for } n = 2 \\ n - 0.5, & \text{for } n > 2 \end{cases} \quad (\text{A.24})$$

$$H = \begin{cases} -0.0870, & \text{for } n = 1 \\ 1.347, & \text{for } n = 2 \\ (n - 0.5)^2 \cdot \left(1 - \frac{2}{(n-0.5)\pi}\right), & \text{for } n \geq 3 \end{cases} \quad (\text{A.25})$$

$$J = \begin{cases} 0.471, & \text{for } n = 1 \\ 3.284, & \text{for } n = 2 \\ (n - 0.5)^2 \cdot \left(1 - \frac{2}{(n-0.5)\pi}\right), & \text{for } n \geq 3 \end{cases} \quad (\text{A.26})$$

A. Appendix to the physical model

clamped/hinged [ch] or [hc]

$$G = n - 0.75, \quad H = J = (n - 0.5)^2 \cdot \left(1 - \frac{2}{(n - 0.5)\pi}\right), \quad \text{for } n \geq 2 \quad (\text{A.27})$$

free/hinged [fh] or [hf]

$$G = \begin{cases} 0, & \text{for } n = 1 \\ n - 0.75, & \text{for } n > 1 \end{cases} \quad (\text{A.28})$$

$$H = \begin{cases} 0, & \text{for } n = 1 \\ (n - 0.75)^2 \cdot \left(1 - \frac{1}{(n - 0.75)\pi}\right), & \text{for } n \geq 2 \end{cases} \quad (\text{A.29})$$

$$J = \begin{cases} \frac{3}{\pi^2}, & \text{for } n = 1 \\ (n - 0.75)^2 \cdot \left(1 - \frac{3}{(n - 0.75)\pi}\right), & \text{for } n \geq 2 \end{cases} \quad (\text{A.30})$$

Bibliography

Warburton, G. B. (June 1954). "The Vibration of Rectangular Plates". In: *Proceedings of the Institution of Mechanical Engineers* 168.1, pp. 371–384. DOI: 10.1243/PIME_PROC_1954_168_040_02.

List of Figures

- 1.1 The interaction loop with auditory feedback. 1
- 1.2 The interaction loop with augmented auditory feedback. 2
- 1.3 Sets of plausible and usable variants of auditory feedback for a specific physical interaction. 3
- 1.4 Two bells and their expected sound. 3
- 1.5 An even, horizontal, rigid, and stationary surface. 4
- 1.6 A person sitting at a table. 5
- 1.7 The roughness of a table may influence, how objects are moved. Sliding on a smooth surface (a), lifting on a rough surface (b). 5
- 1.8 Hierarchical model of processes (activities, actions, and operations) and their corresponding objects (motives, goals, conditions) in activity theory, after Kuutti (1995). 6
- 1.9 Basic structure of an activity, after Kuutti (1995). 7
- 1.10 The basic concept of auditory augmentation, adapted from Bovermann et al. (2010). . . 9
- 1.11 A blank blended sonification diagram (adapted from Tünnermann et al. 2013). 10
- 1.12 Example of a formalized scenario tree created within the comprehension stage of PAM for the scenario “The pack saw the fox. The hounds growled.” 15
- 1.13 Model plausibility rating II depending on the total number of paths P , their average length L , and the amount of non-hypothetical paths N (adapted from Connell and Keane 2006). 15
- 1.14 Plausibility model (adapted from Böhnert and Reszke 2014). 16
- 1.15 The uncanny valley. Adapted from Mori et al. (2012). 17

- 3.1 An undamped harmonic oscillator consisting of a mass m and a spring k 59
- 3.2 The vibration of an undamped harmonic oscillator over time. 60
- 3.3 A damped harmonic oscillator consisting of a mass m , spring k , and damper r 60
- 3.4 The vibration of a damped harmonic oscillator over time. 60
- 3.5 Frequency response of a damped harmonic oscillator. 61
- 3.6 Discretized model of a physical system, described by masses m_i and springs with stiffnesses k_j and dampers r_j . The dampers have been omitted in the drawing for clarity. 62
- 3.7 The normal modes of a vibrating string for different numbers of masses N , leading to N modes. Adapted from Fletcher and Rossing (2010, p. 35). 62
- 3.8 A longish cuboid that is just thin enough to fulfill the requirements of Euler-Bernoulli beam theory for thin bars (relative dimensions: $20 \times 1 \times 1$). 63
- 3.9 Three basic end conditions of a bar (Fletcher and Rossing 2010, p. 61). 63
- 3.10 The mode shapes of the first 7 flexural modes of a thin bar with free boundary conditions (including translation and rotation). 64
- 3.11 A longish cuboid that is just thin enough and wide enough to fulfill the requirements of a thin plate (relative dimensions: $100 \times 20 \times 1$). 64
- 3.12 Thermoelastic weighting factors I_1 of a rectangular metal plate with free boundaries, $\nu=0.3$, and $r_a=2$ 68
- 3.13 Radiation damping α_r and critical frequency f_{cr} of an aluminum plate ($h=3$ mm thick). 68
- 3.14 Relationship between the thermoelastic damping coefficients R_{1t} and c_{1t} for different metals. The values are derived from the basic material constants. 70

List of Figures

3.15	Relationship between thermoelastic damping, thickness, and frequency in case of an aluminum plate.	70
3.16	The decay factors of an aluminum plate due to the three main damping mechanisms: radiation damping, thermoelastic damping, and viscoelastic damping.	70
3.17	Viscoelastic decay factor α_v for glass, as a function of frequency.	71
3.18	Viscoelastic loss factor η_v as a function of longitudinal wave velocity c_L	72
3.19	The overall decay factors of an aluminum plate, in comparison with the individual contributions from radiation damping and viscoelastic damping.	73
3.20	The overall -60 dB decay times of an aluminum plate, in comparison with the individual contributions from radiation damping and viscoelastic damping.	73
3.21	The amplitude weights resulting from excitation of a free rectangular plate at the edge of the long side in the maximum of the $3/0$ mode.	73
3.22	The frequency response of a Hann window, and its approximation by an ideal 3rd-order low-pass filter.	75
3.23	Upper cutoff frequency f_{cH} vs. Brinell hardness HB	75
3.24	The radiation efficiency of an aluminum plate.	77
3.25	Relationship between radiation damping and radiation efficiency of an aluminum plate.	77
3.26	Equivalent parallel model of modal masses m_n , modal stiffnesses k_n , and modal dampings r_n	77
3.27	Comparison between 3 digital resonator implementations at audio rate, based on the simple system from Fig. 3.5: finite difference approximation, simple resonator, Smith-Angell resonator.	79
3.28	Spectrograms (left) and magnitude spectra (right) of synthesized (top) and recorded (bottom) sound of an impacted aluminum plate. Illustration by Czuka (2021).	80
4.1	The concept of AltAR/table.	83
4.2	Simplified block diagram of an auditory augmentation system with structure-borne excitation $e(t)$, transfer function M , and airborne output $s(t)$	83
4.3	Block diagram of the hardware signal flow of the AltAR/table platform.	84
4.4	The AltAR/table hardware platform (top).	84
4.5	The AltAR/table hardware platform (bottom).	85
4.6	Block diagram including transfer functions of the signal paths.	85
4.7	Block diagram including equalization.	86
4.8	Block diagram for input filtering.	86
4.9	Block diagram for output filtering.	86
4.10	Block diagram for impulse response measurements.	86
4.11	Block diagram for the excitation signal, including direct path and input path.	87
4.12	Smoothed magnitude spectra of the structure-borne exciters, measured with airborne microphones.	89
4.13	Smoothed magnitude spectra of the piezos, estimated via sine sweeps from exciters. Plots for the individual piezos, averaged over exciters.	89
4.14	Magnitude spectra of the inverse filters H_{outj} , designed to equalize the frequency response of the structure-borne exciters.	89
4.15	Magnitude spectra of the inverse filters H_{ini} , designed to equalize the frequency response of the contact microphones.	90
4.16	Smoothed magnitude spectra of the signal path between exciters and the 3rd contact microphone (center left).	90
4.17	Graphical user interface for the physical parameters of AltAR/table.	92
4.18	Overall magnitude spectrum of the summed resonators for an aluminum plate.	92
4.19	The principle of resonator H_R and anti-resonator H_R^{-1}	93
4.20	Magnitude responses of a resonator ($f_r = 1$ kHz, $Q = 100$, gain of 40 dB) together with its matched peak EQ, notch EQ (anti-resonator) and the combination of resonator and notch EQ.	93

4.21	Magnitude response of the optimized notch filterbank for the resonator filterbank shown in Fig. 4.18. The grayed-out spectrum shows the raw notch filterbank without gain optimization.	94
4.22	Block diagram including feedback suppression.	94
4.23	Block diagram of the feedback cancellation.	95
4.24	Optical multi-touch frame.	96
4.25	Pencil with infrared-reflective markers.	97
4.26	Coordinate transformations of tracked pencil and table.	97
4.27	Pressure sensing via load cells.	98
5.1	How many bits fit in a rectangular plate?	101
5.2	Measured decay factor in 1/3-octave bands and estimated critical frequency, for a synthesized and real glass plate. Illustration by Czuka (2021).	104
5.3	Measured decay factors α and estimated loss factor η of an impacted plate. Illustration by Czuka (2021).	105
5.4	(a) Cost function with true values (red dot) and paths of the gradient search (white lines). (b) Histogram of the linear assignment cost (dashed line marks best fit). Illustration by Czuka (2021).	106
5.5	Apparatus for experiment 1, including the interface plate, tracked pencil, and MIDI controller (captured via webcam during the experiment).	110
5.6	Experiment software on the screen, in test mode, for experiment 1.	111
5.7	The individual participants' performance: accuracy in parameter identification (length and aspect ratio), accuracy in level identification (3 parameter levels), and accuracy in direction identification (increase and decrease).	112
5.8	Estimated value vs. true value for the 4 combinations of material and parameter.	115
5.9	The number of discriminable levels, plotted against the probability of superiority.	116
5.10	Confusion probability vs. relative area difference for the 36 pairs of plates, based on jitter-corrected true and estimated values. Area differences are relative to the smaller plate's area, respectively; estimated values are rounded to main levels.	116
5.11	Estimated area vs. true area of all stimuli and all participants, (a) for the glass plate and (b) the wooden plate.	117
5.12	Impact patterns of all individual participants (gray = training, black = test), together with the average response time per trial \bar{T}	118
5.13	Frequency ratios, relative to the frequency of mode 2/0 for a quadratic plate, as a function of aspect ratio, for both materials or orthotropy factors.	120
5.14	The adjusted thickness h plotted against the longitudinal wave velocity c_L for all 6 materials used in the experiment.	121
5.15	Frequency factors of the partials of a rectangular plate (relative to the lowest frequency at aspect ratio 1:1) as a function of aspect ratio.	122
5.16	Modal weights of a vibrating bar with free ends as a function of excitation position.	122
5.17	The testing page for a single stimulus.	123
5.18	Individual participants' accuracies together with their average, for the three parameters (pooled over dampings, repetitions, and both other parameters).	124
5.19	The number of discriminable levels of the three parameters (metallicity, material, aspect ratio), plotted against the probability of superiority P_s	128
5.20	Gaussian model for estimating information capacity.	130
5.21	Received vs. sent information for the two dimensions of experiment 1.	130
5.22	Received vs. sent information for the three dimensions of experiment 2.	131
6.1	A black box just appeared in the snow. How would you expect it to sound, if you struck it with a mallet?	133
6.2	The main research question: what makes auditory feedback plausible?	134
6.3	The three sound layers created by Schrödinger's box.	135

List of Figures

6.4	The empty box inside, showing the structure-borne exciters.	136
6.5	The top plate holding the contact microphones, together with its rubber feet that attach it to the rest of the box.	136
6.6	Block diagram of the hardware signal flow of Schrödinger's box.	137
6.7	"Schrödinger's guts": the electronics within Schrödinger's box.	137
6.8	Measured magnitude spectrum of the structure-borne exciters of Schrödinger's box.	138
6.9	Block diagram of the merging strategy of TBOD and FBOD.	139
6.10	Block diagram of time-based onset detection (TBOD).	139
6.11	Block diagram of frequency-based onset detection (FBOD).	140
7.1	The concept of the Science by Ear workshop series: sonification experts, artists, and domain scientists forming interdisciplinary teams for quick prototyping.	145
7.2	Wall plug and wireless USB dongle for real-time electric power measurement.	146
7.3	The BRIX ₂ physical computing and sensor platform. Photo: Sebastian Zehe, 2014.	147
7.4	The table platform in SBE4, as seen from above.	147
7.5	The IEM Cube, serving as the room platform in SBE4. Photo: W. Hummer / KUG.	147
7.6	Writing resonances.	148
7.7	Exploration table.	148
7.8	Spatial partitioning of the exploration table interface.	148
7.9	Sonic floor plan.	149
7.10	Smart kettle.	149
7.11	Standby door.	150
7.12	3D gestural mouse.	150
7.13	Hob assistant.	151
7.14	Interleave.	151
7.15	Sound-centered auditory augmentation. The task produces a sound (auditory feedback) that is modulated by the data.	153
7.16	Object-centered auditory augmentation. The object itself is (metaphorically) modulated by the data.	154
7.17	Task-centered auditory augmentation. The task produces data that is used to generate sound.	154
7.18	Data-centered auditory augmentation. Data generates sound that is modulated by the given task.	154
7.19	Block diagram of auditory augmentation.	154
8.1	Case studies of auditory augmentations: table, room, and human body.	159
8.2	Concept of the Mondrian Table, inspired by Piet Mondrian's "Composition II in Red, Blue, and Yellow" (1930).	160
8.3	Blended sonification diagram of the Mondrian Table.	160
8.4	The Mondrian graphic tablet. Microphone positions are marked by a red 'M'.	161
8.5	Example image, generated with the Mondrian Generator, serving as test data.	162
8.6	The auditory coloring book.	163
8.7	Drawings from individual participants, overlain by the corresponding "correct answer".	164
8.8	The kitchen of the Institute of Electronic Music and Acoustics (IEM). Photos: Till Bovermann / IEM.	165
8.9	One week of electric power consumption during the second evaluation study, 12–18 March 2018.	167
8.10	Average ratings and 95 % confidence intervals across participants, for the dimensions of valence, arousal, and dominance, respectively.	170
8.11	The GUI of the smartphone app during the three phases of the experiment: (a) training I, (b) training II, (c) test.	173
8.12	Histogram of scores from the multiple-choice questionnaire, for EG and CG.	174
8.13	Experiment data of participant 16.	175

8.14	Number of discriminable levels for different thresholds d_t (Tab. 8.2) and their corresponding probability of superiority (P_s). (a) depicts the number of levels as a function of P_s . (b) shows the same information for 4 selected values of P_s , together with the standard deviation across participants.	176
8.15	Estimated vs. real reverberation, pooled over all participants, split into upward and downward level changes.	176
8.16	Average rating and standard deviation for the questions on the participants' impressions.	177
8.17	Wiggers diagram. Adapted from Wikimedia Commons (users adh30, DanielChangMD, DestinyQx, and xavax). Licensed under CC BY-SA 4.0.	179
8.18	The system for synchronous signal acquisition of ECG and PCG through the same audio interface.	180
9.1	Someone shaking a box to guess its contents from the resulting sound. A task we want to facilitate.	185
9.2	The principle of enhanced inter-stimulus contrast by enhanced intra-stimulus contrast.	186
9.3	Overall block diagram of real-time ACE.	187
9.4	Principle of dynamics expansion.	188
9.5	Principle of lateral inhibition.	188
9.6	The photo of a white duck in three versions, and the drawing of a famous cartoon duck.	188
9.7	Simplified block diagram of one channel of sub-band processing.	189
9.8	Overall block diagram of spectral ACE	189
9.9	Detailed block diagram of one channel of sub-band processing (SP)	190
9.10	Spectrograms of a test sound in 4 conditions: (a) original recording, (b) with spectral sharpening, (c) with spectral sharpening and spectral dynamics expansion, (d) with spectral sharpening, spectral dynamics expansion, and decay prolongation.	192
9.11	Principle of decay prolongation.	193
9.12	Spectrograms of a test signal without (a) and with (b) vibrato expansion ($\lambda=0.05$).	194
9.13	Signal waveform without (left, Snd. S2.1) and with temporal ACE (right, Snd. S2.3).	195
9.14	Experiment results. Median plausibility of the 10 videos and their two main clusters, for all 26 combinations of spectral dynamics (ρ/β) and decay time T_{60} at 1 kHz, respectively. Values reach from 0 (very implausible) to 1 (very plausible).	198
9.15	Schematic drawing of the proposed ACE hear-through system.	200
9.16	Mobile ACE based on Bela.	201
















List of Tables

2.1	Overview of listening modes (adapted from Supper and Bijsterveld 2015).	25
3.1	Basic material constants of isotropic and orthotropic plates.	65
3.2	Material constants for radiation damping.	69
3.3	Material constants for thermoelastic damping.	69
3.4	Material constants for viscoelastic damping.	71
4.1	Coordinates of inputs (contact microphones) and outputs (structure-borne exciters), relative to the upper left corner, in mm.	90
5.1	True and estimated loss factors η_v .	107
5.2	True and estimated values for critical frequency f_{cr} , frequency factor Φ , aspect ratio r_a , and area S , based on measurements on synthesized and real physical plates.	108
5.3	Model coefficients of the rendered plates.	110
5.4	Levels of the independent parameters length and aspect ratio.	110
5.5	Confusion matrix for the parameter dimension of a change, for both materials separately, pooled over participants.	113
5.6	Confusion matrix for the 3 main levels of length, for both materials separately, pooled over participants.	113
5.7	Confusion matrix for the 3 main levels of aspect ratio, for both materials separately, pooled over participants.	113
5.8	Confusion matrix for the sign of a change in length or aspect ratio (+ increase, – decrease), for both materials separately, pooled over participants.	114
5.9	Cohen's d between neighboring main levels of length and aspect ratio, pooled over participants, for both materials, respectively.	115
5.10	Top 10 confused pairs of plates, together with their probability of confusion.	116
5.11	Surface areas of the rendered plates in m^2 for all combinations of the main levels of length and aspect ratio.	117
5.12	Model coefficients of the rendered plates. The thermoelastic constants of metal material are equally used for non-metals.	122
5.13	Top 16 confused pairs of plates, together with their probability of confusion.	125
5.14	Confusion matrix for metallicity, for both damping conditions separately, pooled over participants, rigidities and aspect ratios.	125
5.15	Confusion matrix for metallicity, for rigidities separately, pooled over participants, dampings, and aspect ratios.	126
5.16	Confusion matrix for metallicity, for aspect ratios separately, pooled over participants, dampings, and rigidities.	126
5.17	Confusion between rigidities, for both dampings separately, pooled over participants, metallicities, and aspect ratios.	126
5.18	Confusion between rigidities for non-metals and metals separately, pooled over participants, dampings, and aspect ratios.	126
5.19	Confusion between rigidities for the three aspect ratios, pooled over participants, dampings, and metallicities.	127

List of Tables

5.20	Confusion between aspect ratios, for both dampings separately, pooled over participants, metallicities, and rigidities.	127
5.21	Confusion between aspect ratios, dependent of rigidity, pooled over participants, metallicities, and dampings.	127
8.1	Physical properties of the reference models.	163
8.2	Probability of superiority P_s corresponding to the thresholds d_t and results for the number of levels discriminable by all, CF, and SF listeners.	176
9.1	Properties of the audio/video files used in the experiment: material and object category, original filename, starting time in the recording in s, and acoustical descriptors, i.e., spectral centroid (SCG), spectral bandwidth (SB), and average -60 dB decay time T_{60}	197

List of Supplementary Material

4.1 Python script for forwarding multitouch to OSC. </> https://github.com/m---w/multitouch2osc	97
4.2 Source code of USB MIDI balance. </> https://github.com/m---w/HX711_MIDI	98
4.3 Video demonstration of AltAR/table.  https://phaidra.kug.ac.at/o:126460	98
4.4 Source code of AltAR/table. </> https://github.com/m---w/altar	98
5.1 Source code of the experiment (incl. sound generation, procedure, media files). </>  https://github.com/m---w/experiment_listening_to_rectangular_plates	123
6.1 Demo videos of Schrödinger's box.  https://phaidra.kug.ac.at/o:126434	141
6.2 Source code of Schrödinger's box. </> https://github.com/m---w/schroedingers-box	141
7.1 Video demo of Writing Resonances.  https://phaidra.kug.ac.at/o:126385	148
7.2 Video demo of Exploration Table.  https://phaidra.kug.ac.at/o:126386	148
7.3 Video demo of Smart Kettle.  https://phaidra.kug.ac.at/o:126384	149
7.4 Video demo of Standby Door.  https://phaidra.kug.ac.at/o:126382	150
7.5 Video demo of 3D Gestural Mouse.  https://phaidra.kug.ac.at/o:126383	150
7.6 Video demo of Hob Assistant.  https://phaidra.kug.ac.at/o:126381	151
7.7 Video demo of Interleave.  https://phaidra.kug.ac.at/o:126380	152
8.1 Video demo of the Mondrian Table.  https://phaidra.kug.ac.at/o:69732	163
8.2 Video of the augmented kitchen.  https://dx.doi.org/10.1162/comj_a_00553	168
8.3 Parameter settings for the RadioReverb experiment. </> https://dx.doi.org/10.1162/comj_a_00553	172
8.4 Sound examples for CardioScope.  http://dx.doi.org/10.4119/unibi/2938001	181
9.1 Sound examples for real-time auditory contrast enhancement (ACE).  https://doi.org/10.4119/unibi/2935786	186
9.2 Video demos of the ACE hear-through system.  https://phaidra.kug.ac.at/o:91924	201
9.3 Source code of ACE. </> https://git.iem.at/weger/ace	201

Related publications

The following publications appeared in direct connection with my doctoral research. Some of them form the basis of this dissertation.

- Aldana Blanco, Andrea Lorena, Marian Weger, Stefan Grautoff, Robert Höldrich, and Thomas Hermann (2019). "CardioScope: ECG sonification and auditory augmentation of heart sounds to support cardiac diagnostic and monitoring." In: *Interactive Sonification Workshop (ISon)*. Stockholm, Sweden, pp. 115–122.
- Czuka, Martin, Marian Weger und Robert Höldrich (2021). „Klangsynthese und akustische Erkennung rechteckiger Platten". In: *DAGA - Jahrestagung für Akustik*. Vienna, Austria.
- Groß-Vogt, Katharina, Marian Weger, and Robert Höldrich (2018). "Exploration of Auditory Augmentation in an Interdisciplinary Prototyping Workshop". In: *Forum Media Technology*. St. Pölten, Austria, pp. 10–16.
- Groß-Vogt, Katharina, Marian Weger, Robert Höldrich, Thomas Hermann, Till Bovermann, and Stefan Reichmann (June 2018). "Augmentation of an Institute's Kitchen: An Ambient Auditory Display of Electric Power Consumption". In: *International Conference on Auditory Display (ICAD)*. Houghton, Michigan, pp. 105–112. DOI: 10.21785/icad2018.027.
- Groß-Vogt, Katharina, Marian Weger, Matthias Frank, and Robert Höldrich (Apr. 5, 2021). "Peripheral Sonification by Means of Virtual Room Acoustics". In: *Computer Music Journal* 44.1, pp. 71–88. DOI: 10.1162/comj_a_00553.
- Hermann, Thomas and Marian Weger (June 2019). "Data-driven Auditory Contrast Enhancement for Everyday Sounds and Sonifications". In: *International Conference on Auditory Display (ICAD)*. Newcastle upon Tyne, UK, pp. 83–90. DOI: 10.21785/icad2019.005.
- Weger, Marian, Michael Aurenhammer, Thomas Hermann, and Robert Höldrich (2022). "The information capacity of plausible auditory augmentations: percussion of rectangular plates." In: *Interactive Sonification Workshop (ISon)*. Delmenhorst, Germany.
- Weger, Marian, Thomas Hermann, and Robert Höldrich (June 2018). "Plausible Auditory Augmentation of Physical Interaction". In: *International Conference on Auditory Display (ICAD)*. Houghton, Michigan, pp. 97–104. DOI: 10.21785/icad2018.024.
- (June 2019). "Real-time Auditory Contrast Enhancement". In: *International Conference on Auditory Display (ICAD)*. Newcastle upon Tyne, UK, pp. 254–261. DOI: 10.21785/icad2019.026.
- (2022). "AltAR/table: a platform for plausible auditory augmentation." In: *International Conference on Auditory Display (ICAD)*. Virtual Conference.
- Weger, Marian and Robert Höldrich (2019). "A hear-through system for plausible auditory contrast enhancement". In: *Proceedings of Audio Mostly*. Nottingham, UK: ACM, pp. 1–8. DOI: 10.1145/3356590.3356593.
- Weger, Marian, Iason Svoronos-Kanavas, and Robert Höldrich (2022). "Schrödinger's box: an artifact to study the limits of plausibility in auditory augmentations." In: *Audio Mostly*. St. Pölten, Austria: ACM.