EAA Euroregio - EAA Winter School
18-21 March 2013 in Merano

AIA-DAGA 2013 – Conference on Acoustics including the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA)

Hot Topics in Acoustics:

Cutting Edge in Spatial Audio

Lecture material by:

Ville Pulkki
Piotr Majdak
Craig Jin
Sascha Spors
Florian Völk

Course editor:
Franz Zotter

Coordinators:
Kristian Jamrosic
Luigi Maffei
Michael Vorländer

online publication by EAA Documenta Acustica

Abstract. Admitted: spatial audio is not an entirely new subject. Nevertheless research has been bringing forward big leaps and cutting-edge technology in spatial audio for the entire last decade, with many excellent experts contributing to research and development. We can bluntly say that today our understanding of spatial audio is unprecedented in many ways: We know much more about neural mechanisms of spatial hearing, experimental data are available that cover many relationships in great precision, and models of spatial hearing become more precise and applicable. On the other hand, we can demonstrate various high-quality sound reinforcement systems showing the power of binaural (headphone-based) or loudspeaker-based holophonic technologies, such as wave field synthesis and Ambisonics, or parametric audio coding methods exploiting psychoacoustic effects. Virtual acoustics rendering systems with room auralization are complemented by recording concepts such as spherical arrays with improved spatial resolution, distributed intelligent array technologies for audio scene transcription, and parametric audio coding for first order microphone arrays. The audible result is finally most relevant: profound evaluation of the various technical methods is currently being researched. It reveals where methods are most effective, and which combinations of technology could yield our research into an interesting future.

Preface. Michael Vorländer asked me in 2012 whether I could imagine editing a two-day program for the EAA Winter School courses in Meran 2013. After accepting this, I found three pleasing surprises: (1) The great interest of lecturers to come, present, and show demo material, (2) the registration for spatial audio closing four months earlier because of the many interested participants, and (3) the effort of my colleagues Matthias Frank and Hannes Pomberger to finally drive a 23 channel loudspeaker system from Graz to Meran with me to support practical demos.

There were about 50 participants, most of whom PhD students from various countries (Germany, Italy, Austria, Denmark, Poland, France, Spain, Croatia, Slovenia, ...). The lecture program was structured by demos, introduction rounds, and discussion, between the announced program points:

- Ville Pulkki: Physiology of binaural hearing and basics of time-frequency-domain spatial audio processing
- Piotr Majdak: Sound localization in sagittal planes
- Craig Jin: Super-resolution sound field analyses
- Maurizio Omologo: Sound field analysis using distributed microphone array networks
- Sascha Spors: Spatial sound synthesis with loudspeakers
- Hagen Wierstorf: Virtual source localization in wave field synthesis
- Florian Völk: Experiments with loudspeaker-based virtual acoustics
- Matthias Frank: Source width of horizontal amplitude panning

Lectures by Maurizio Omologo, Matthias Frank, and Hagen Wierstorf were held in the Winter School but are not part of this booklet. These lecturers were included in the course because of their relevance, despite I knew that these busy people were not be able to contribute a written contribution, this time.

Franz Zotter, June 20th, Graz, 2013.

Contents

Physiology of binaural hearing and Basics of time-frequency-domain spatial audio processing

Ville Pulkki
Department of signal processing and acoustics
Aalto University
POBox 13000, 00076 Aalto
ville.pulkki@aalto.fi

Learning material for EAA Winter School 2013

January 29, 2013

1 Mammalian auditory pathway

The pathway of the sound, and the neural impulses evoked by it, is reviewed here starting from the ear and continuing up to the point at which the neural signals from the eyes and the ears meet, as shown in Fig. 1. The sound received by the ear travels through the outer ear and the middle ear and arrives at the inner ear. There, the hair cells in the cochlea transform the sound into neural impulses, which are divided into frequency bands. Then, the signal traverses via the auditory nerve into the cochlear nucleus (CN), which is located in the brainstem [1]. The various cells in the CN send different responses to different targets in the auditory system. In both hemispheres, the CN projects temporally accurate responses via the dorsal stream into the MSO and the LSO. The CN also projects responses directly into the inferior colliculus (IC) via the ventral stream. The ventral and dorsal streams may be considered as the origins of the *what* and *where* streams, respectively, reflecting the division of the *what* and *where* processing streams of auditory cortical processing [2]. Furthermore, the auditory information included in the sound spectrum is thought to be analyzed in the *what* stream, whereas the spatial information of different sound events in

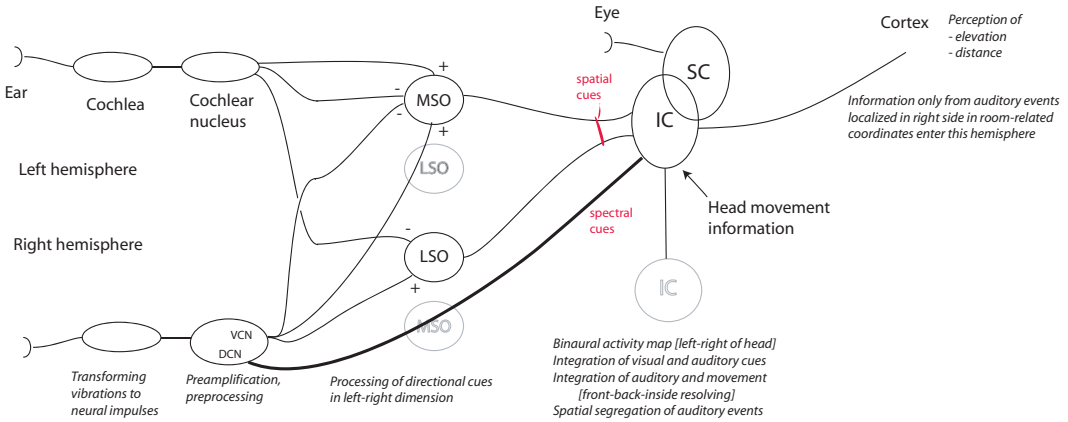

Figure 1: Schematic drawing of auditory pathway with organs devoted to spatial hearing

the auditory scene is thought to be analyzed in the *where* stream. The MSO and the LSO play an important role in the localization and spatial hearing due to the fact that the activity originating from the two ears converges for the first time in the MSO and LSO, and they are known to be sensitive to differences in binaural signals [3].

The MSO receives both excitation and inhibition from the CN of both hemispheres [4], and its cells are known to be sensitive to interaural time difference (ITD) [5]. In [5] the functionality of the MSO neurons is described as coincidence counters that respond to a specific interaural phase difference (IPD) in such a manner that most of the neurons sharing the same characteristic frequency are sensitive to an IPD of $\pi/4$ at low frequencies. The output of the MSO is delivered mainly to the ipsilateral IC [6]. The LSO receives excitation from the ipsilateral CN and inhibition from the contralateral CN [7], and the sensitivity of the LSO to interaural level difference (ILD) has been shown by [8]. There is evidence that the LSO neurons act as phase-locked subtractors that can respond to very fast changes in the input signals, with an integration time of as low as 2 ms [9]. The excitatory output of the LSO is routed to the IC in the contralateral hemisphere.

The role of the IC in binaural processing is still somewhat unclear, despite the numerous studies that have measured the responses of IC neurons [10]. This may be related to the versatility of the IC. What is known of the IC is that it transmits the spatial information from the CN, MSO, and LSO to the auditory cortex and the SC, and it may modify the information in the process [10].

The superior colliculus (SC) is located next to the IC, and it has multiple layers in the nucleus, including layers, for example, for visual information as well as for sound [11, 12]. The SC has been found to be one of the organs responsible for cross-modal interaction, and it is also involved in steering the focus of attention towards the stimuli [13, 14]. Interestingly, topographical organization of the auditory space has been found in the SC [12, 1]. The SC includes neurons that respond to multimodal stimulation originating from the same spatial location [11, 15] and also a map of the auditory space that is aligned with the visual map in such a manner that the neurons responsive to auditory or visual stimuli from a certain direction can be found close to one another [11, 12].

2 Basics of binaural psychoacoustics

The spatial hearing of human listeners has been studied actively in psychoacoustical experiments over the years [16]. The review presented in this section is limited to the most relevant aspects pertaining to this study.

The sound localization of humans is based on binaural cues between the ear canal signals resulting from differences in the path of the sound from the sound source to the two ears and on spectral cues caused by the reflections of the sound on the pinna and torso. The binaural cues consist of ITD, ILD, and envelope time-shifts [16]. Typically, the auditory system can use all of the aforementioned cues in the localization process since they all point in the same direction when a single plane wave arrives at the ears of a listener. However, listening tests conducted in controlled scenarios have demonstrated that a modification of even one of the binaural differences away from a zero value is sufficient to shift the perceived lateral position of the evoked auditory image away from the center [17, 18, 19], for example by modifying the stimulus so that only the ILD is non-zero. This tendency of the auditory system to favor the off-median plane cues in localization is called the *lateral preference* principle in this article. In the case of conflicting non-zero binaural cues, the ITD has been found to be the dominant cue for both broadband and low-pass filtered stimuli, whereas the ILD has been found to dominate the localization with high-pass filtered stimuli [20, 21].

In free-field listening with only one point-like sound source, the sound emitted is perceived as a narrow auditory image [16]. The accuracy of the localization in such conditions has been found

to depend on the direction of arrival, the type, and the length of the sound signal [22, 23, 24]. The localization accuracy has often been measured as the minimum audible angle, i.e. the angle the sound source needs to be shifted from its original direction before the subject can detect the change [25]. This resolution has been found to be approximately $\pm 1^\circ$ directly in front and to decrease gradually to approximately $\pm 10^\circ$ when the sound is moved to the side on the horizontal plane [26, 16].

The spatial perception becomes more challenging when the sound heard by the listener actually consists of an ensemble of independent signals emitted by multiple sound sources around the listener. If the task of the listener is to localize a particular sound event from the ensemble, the ensemble can be thought to consist of a target sound and distracter(s) that hinder the localization task. Furthermore, the reflections of the target sound in a reverberant environment can be considered as distracters as well, and experiments on the precedence effect [27] have shown that in such environments, listeners perceive that the sound is emitted from the direction of the direct sound. The amount of decrease in the localization of the target sound caused by the distracter(s) has been found to be dependent on several factors, such as the number of distracters, the signal types, the frequency contents, the signal to noise ratio, and the onset and offset times of the target and the distracter(s) [28, 29, 30, 31, 32, 16, 27, 33]. Moreover, in conditions that consist of independent noise bursts, the length of the simultaneous noise bursts emitted has been found to have an effect on whether the ensemble is perceived as point-like or wide [34]. Additionally, the perceived width of an ensemble emitting incoherent noise has been found to be slightly narrower than the loudspeaker span employed in the reproduction, and that the ends of the distributed ensemble are perceived relatively accurately whereas the center area is perceived less clearly [35].

If the ensemble consists of multiple speech sources, the scenario is related to the "cocktail party effect" [36], where it has been shown that listeners are able to segregate speech in multi-talker situations and are able to both localize the different speakers and to identify the sentences spoken by the different speakers, although the speech intelligibility and the localization accuracy are both worse than in single-source scenarios [29, 37, 16, 38].

It should be noted that the overall perception of the auditory scene is a result of a cognitive process affected not only by the auditory information, but also by the visual information and the head movements. Support for such an interaction can be found in the experiments of the ventriloquism [39, 40] and the McGurk [41] effects. The former of the mentioned effects demonstrates a shift in the localization of a sound towards the visual image of the sound source, and the latter demonstrates the change in the heard utterances caused by conflicting visual and auditory cues. As mentioned above in Sec. 1, the cortical pathways from the eyes and the ears meet in the SC. Therefore, the aforementioned effects may be partly explained by the processing in the SC.

3 Basics of time-frequency-domain spatial audio processing

The spatial properties of sound perceivable by humans are the directions and distances of sound sources in three dimensions, and the effect of the room on sound. In addition, the spatial arrangement of sound sources affects the timbre, which corresponds to perceived sound color. The directional resolution of spatial hearing is limited within auditory frequency bands [16]. In principle, all sound within one critical band can be perceived only as a single source with broader or narrower extent. In some special cases, binaural narrow-band sound stimulus can be perceived as two distinct auditory objects, but the perception of three or more concurrent sources is generally not possible, which differs from visual perception, where already one eye can detect the directions of numerous visual objects sharing the same color.

The limitations of spatial auditory perception imply that such spatial realism that is needed in visual reproduction is not needed in audio. In other words, the spatial accuracy in reproduction of

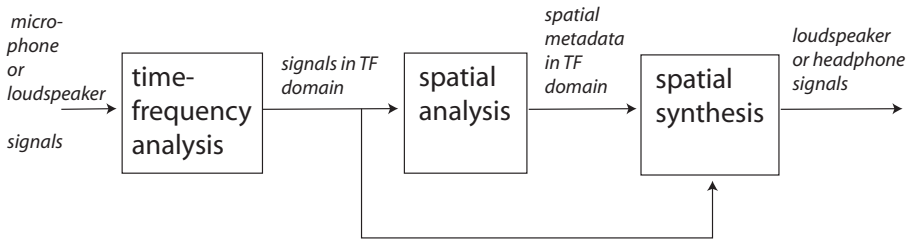

Figure 2: Basic principle of parametric spatial sound reproduction.

acoustical wave field can be compromised without decreasing perceptual quality. This relaxation of boundary conditions has been recently taken into use in spatial audio.

The basic paradigm is shown in Fig. 2. The input is obtained using multiple microphones, the signals of which are divided into a time-frequency presentation. An analysis of spatial properties of the sound field is performed on the microphone signals, and finally the synthesis of sound is based both on the audio signal(s) and on parametric metadata obtained from the analysis. The number of audio signals can be one or larger, and the signal(s) can be derived either by summing some microphone signals to a downmix signal, or by transmitting all of the signals intact.

3.0.1 Directional audio coding

In directional audio coding (DirAC) [42, 43], it is assumed that at one time instant and at one critical band the spatial resolution of auditory system is limited to decoding one cue for direction and another for inter-aural coherence. It is further assumed, that if the direction and diffuseness of sound field is measured and reproduced correctly, a human listener will perceive the directional and coherence cues correctly.

In the analysis phase, the direction and diffuseness of the sound field is estimated in auditory frequency bands depending on time, forming the metadata transmitted together with a few audio channels. DirAC thus assumes that the field consists of a single incoming plane wave superposed with a perfectly diffuse field at one frequency band. Although the direction and diffuseness could be measured with many techniques, most of the implementations of DirAC utilize the B-format input followed by some kind of energetic analysis of sound field.

The target of directional analysis, which is shown in Fig. 3, is to estimate at each frequency band the direction of arrival of sound, together with an estimate if the sound is arriving from one or multiple directions simultaneously. In principle, this can be performed using several techniques, however, the energetic analysis of sound field has been found to be suitable, as shown in Fig. 3. The energetic analysis can be performed when the pressure signal and velocity signals in 1–3 dimensions are captured from a single position. In first-order B-format signals, the omnidirectional signal is called the W-signal, which has been scaled down by $\sqrt{2}$. The sound pressure can be estimated as $P = \sqrt{2}W$, expressed in STFT domain. The X-, Y- and Z-channels have the directional pattern of a dipole directed along the Cartesian axis, which together form a vector $\mathbf{U} = [X, Y, Z]$. The vector estimates the sound field velocity vector; it is also expressed in the STFT domain. The energy E of the sound field can be computed as

$$E = \frac{\rho_0}{4}||\mathbf{U}||^2 + \frac{1}{4\rho_0 c^2}|P|^2, \tag{1}$$

where ρ_0 stands for the mean density of air, and c signifies the speed of sound. The capturing of B-format signals can be achieved with either coincident positioning of directional microphones, or with closely spaced set of omnidirectional microphones. In some applications, the microphone

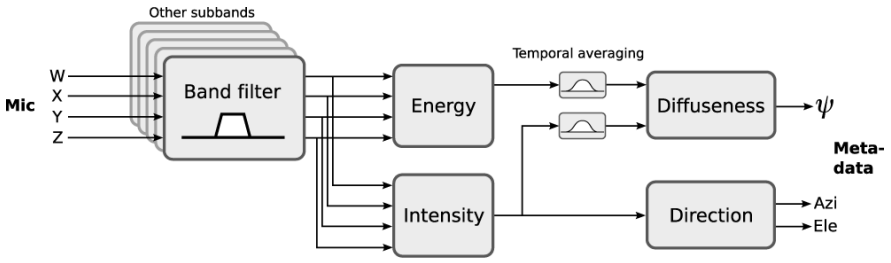

Figure 3: DirAC analysis.

signals might be formed in the computational domain, i.e. simulated. The analysis is repeated as frequently as is needed for the application, typically with the update frequency of 100–1000 Hz.

The intensity vector $\mathbf{I}$ expresses the net flow of sound energy as a 3D vector. It can be computed as

$$\mathbf{I} = \overline{P}\mathbf{U}, \tag{2}$$

where $\overline{(\cdot)}$ denotes complex conjugation. The direction of sound is defined as the opposite direction of the intensity vector at each frequency band. The direction is denoted as corresponding to angular azimuth and elevation values in the transmitted metadata. The diffuseness of sound field is computed as

$$\psi = 1 - \frac{\|\mathrm{E}\{\mathbf{I}\}\|}{c\mathrm{E}\{E\}}, \tag{3}$$

where E is the expectation operator. The outcome of this equation is a real-valued number between zero and one, characterizing whether the sound energy is arriving from a single direction, or from all directions. This equation is appropriate in the case in which the full 3D velocity information is available.

In the "low-bitrate" version of DirAC, only one channel of audio is transmitted. The audio channel may also be further compressed to obtain lower transmission data rate. The version with more channels is shown as "high-quality version", where the number of transmitted channels is three for horizontal reproduction, and four for 3D reproduction. In high-quality version the analysis may be conducted in the receiving end.

In the case of low-bitrate version, the single transmitted channel is divided into diffuse and non-diffuse streams. Non-diffuse stream is reproduced over amplitude panning, and diffuse stream by applying the same signal to all loudspeakers with decorrelation. The method produces good quality for telecommunication purposes. However, the quality of multiple simultaneous sources and reverberation is degraded. With high-quality version, more channels are transmitted, and the audio signals for the loudspeakers are derived by matrixing followed by slight decorrelation. It has been proven that high-quality version of DirAC produces better perceptual quality in loudspeaker listening than other available techniques using the same microphone input [44].

The high-quality version of DirAC synthesis, shown in Fig. 4, receives all B-format signals, from which a virtual microphone signal is computed for each loudspeaker direction. The used directional pattern is typically a dipole. The virtual microphone signals are then modified in nonlinear fashion, depending on the metadata. The low-bit-rate version of DirAC is not shown in the figure, however; in it only one channel of audio is transmitted. The difference in processing is that all virtual microphone signals would be replaced by the single channel of audio received. The virtual microphone signals are divided into two streams (the diffuse and the non-diffuse streams), which are processed separately.

The non-diffuse stream is reproduced as point sources using vector base amplitude panning (VBAP) [45]. In panning, a monophonic sound signal is applied to a subset of loudspeakers

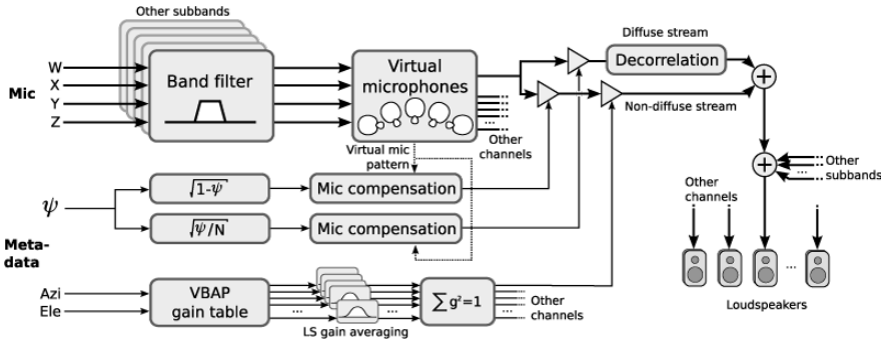

Figure 4: DirAC synthesis.

after multiplication with loudspeaker-specific gain factors. The gain factors are computed using the information of the loudspeaker setup and the specified panning direction. In the low-bit-rate version, the input signal is simply panned to the directions implied by the metadata. In the high-quality version, each virtual microphone signal is multiplied with the corresponding gain factor.

In many cases, the direction in metadata is subject to abrupt temporal changes. To avoid artifacts, the gain factors for loudspeakers computed with VBAP are smoothed by energy-weighted temporal integration with frequency-dependent time constant equaling about 50 cycle periods at each band, which effectively removes the artifacts. However, the changes in direction are not perceived to be slower than without averaging in most cases.

The aim of the synthesis of **the diffuse stream** is to create perception of sound that surrounds the listener. In the low-bit-rate version, the diffuse stream is reproduced by decorrelating the input signal and reproducing it from every loudspeaker. In the high-quality version, the virtual microphone signals of diffuse streams are already incoherent to some degree. They must be decorrelated only mildly. This approach provides better spatial quality for surrounding reverberation and ambient sound than the low-bit-rate version does.

Prior applications concentrated on cases in which the directional properties of sound are captured with real microphones from a live situation. It is also possible to use DirAC in virtual or mixed realities [46]. In these applications, the directional metadata connected to a DirAC stream are defined by the user. For example, a single channel of audio is spatialized as a point-like virtual source with DirAC when the same direction for all frequency channels is added as metadata to the signal. In some cases, it would be beneficial to control the perceived extent or width of the sound source. A simple and effective method for this is to use a different direction value for each frequency band, where the values are distributed inside the desired directional extent of the virtual source. This is effective especially with signals with temporally relatively smooth envelope, such as background ambient sounds or noisy signals.

References

[1] A. G. Møller, Ed., *Hearing: Anatomy, Physiology, and Disorders of the Auditory System*, pp. 75– 150, Academic Press, San Diego, CA, 2nd edition, 2006.

[2] J. P. Rauschecker and B. Tian, "Mechanisms and streams for processing of ÔÔwhatÕÕ and ÔÔwhereÕÕ in auditory cortex," *Proc. of the Natl. Acad. Sci. U.S.A*, vol. 97, no. 22, pp. 11800 – 11806, Oct. 2000.

[3] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiol. Rev.*, vol. 90, no. 3, pp. 983 – 1012, Jul. 2010.

[4] N. B. Cant and R. L. Hyson, "Projections from the lateral nucleus of the trapezoid body to the medial superior olivary nucleus in the gerbil," *Hear. Res.*, vol. 58, no. 1, pp. 26–34, 1992.

[5] B. Grothe, "Sensory systems: New roles for synaptic inhibition in sound localization," *Nat. Rev. Neurosci.*, vol. 4, pp. 540–550, 2003.

[6] T. Yin, "Neural mechanisms of encoding binaural localization cues in the auditory brainstem," in *Integrative functions in the mammalian auditory pathway*, D. Oertel, A. Popper, and R. Fay, Eds., pp. 99–159. Springer, New York, 2002.

[7] D. H. Sanes, "An in vitro analysis of sound localization mechanisms in the gerbil lateral superior olive," *J. Neuroscience*, vol. 10, no. 11, pp. 3494–3506, 1990.

[8] D. J. Tollin, K. Koka, and J. J. Tsai, "Interaural level difference discrimination thresholds for single neurons in the lateral superior olive," *J. Neuroscience*, vol. 28, no. 19, pp. 4848–4860, 2008.

[9] P. Joris, "Envelope coding in the lateral superior olive. II. Characteristic delays and comparison with responses in the medial superior olive," *J. Neurophysiol*, vol. 76, pp. 2137–2156, Oct. 1996.

[10] D. Irvine, "Physiology of the auditory brainstem," in *The Mammalian Auditory Pathway: Neurophysiology*, A. N. Popper and R. R. Fay, Eds., pp. 157 – 231. Springer-Verlag, New York, NY, USA, 1992.

[11] B. Gordon, "Receptive fields in deep layers of cat superior colliculus," *J. Neurophysiol.*, vol. 36, no. 2, pp. 157–178, Mar. 1973.

[12] A. R. Palmer and A. J. King, "The representation of auditory space in the mammalian superior colliculus," *Nature*, vol. 299, pp. 248–249, Sept. 1982.

[13] B. E. Stein and M. A. Meredith, *The Merging of the Senses*, MIT Press, Cambridge, MA, USA, 1993.

[14] G. A. Calvert, "Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies," *Cereb. Cortex*, vol. 11, no. 12, pp. 1110 – 1123, Dec. 2001.

[15] C. K. Peck, "Visual-auditory interactions in cat superior colliculus: their role in the control of gaze," *Brain Res.*, vol. 420, no. 1, pp. 162 – 166, Sep. 1987.

[16] J. Blauert, *Spatial Hearing. The psychophysics of human sound localization*, pp. 37 – 50, 140 – 155, 164 – 176, MIT Press, Cambridge, MA, USA, 2nd edition, 1997.

[17] Lord Rayleigh, "On our perception of sound direction," *Phil. Mag. Series 6*, vol. 13, no. 74, pp. 214 – 232, 1907.

[18] G. von Békésy, "Zur Theorie des Hörens. Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleighheit der beiderseitigen Schalleinwirkungen," *Physik. Zeitschr.*, pp. 824–835, 857–868, 1930.

[19] W. A. Yost, "Lateral position of sinusoids presented with interaural intensive and temporal differences," *J. Acoust. Soc.Am.*, vol. 70, no. 2, pp. 397–409, Aug. 1981.

[20] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648 – 1661, Mar. 1992.

[21] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2219 – 2236, May 2002.

[22] S. S. Stevens and E. B. Newman, "The Localization of Actual Sources of Sound," *Am. J. Psychol.*, vol. 48, no. 2, pp. 297 – 306, Apr. 1936.

[23] G. Boerger, *Die Lokalisation von Gausstönen*, Ph.D. thesis, Technische Universität, Berlin, Germany, 1965.

[24] M. B. Gardner, "Lateral localization of 0° or near-0° oriented speech signals in anechoic conditions," *J. Acoust. Soc. Am.*, vol. 44, no. 3, pp. 797 – 802, 1968.

[25] A. W. Mills, "On the minimum audible angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237 – 246, 1958.

[26] R. Preibisch-Effenberger, *Die Schallokalisationsfähigkeit des Menschen und ihre audiometrische Verwendung zur klinischen Diagnostik*, Ph.D. thesis, Technische Universität, Dresden, Germany, 1965.

[27] R. Litovsky, S. Colburn, W. A. Yost, and S. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633 – 1654, Oct. 1999.

[28] J. L. Flanagan and B. J. Watson, "Binaural unmasking of complex signals," *J. Acoust. Soc. Am.*, vol. 40, no. 2, pp. 546 – 468, 1966.

[29] R. T. Carhart, T. W. Tillman, and E. S. Greetis, "Perceptual masking in multiple sound backgrounds," *J. Acoust. Soc. Am.*, vol. 45, no. 3, pp. 694 – 703, 1969.

[30] D. McFadden and E. G. Pasanen, "Lateralization at high frequencies based on interaural time differences," *J. Acoust. Soc. Am.*, vol. 59, no. 3, pp. 634 – 639, Mar. 1976.

[31] A. Kohlrausch, "The influence of signal duration, signal frequency and masker duration on binaural masking level differences," *Hear. Res.*, vol. 23, no. 3, pp. 267 –273, Feb. 1986.

[32] B. Kollmeier and R. H. Gilkey, "Binaural forward and backward masking: Evidence for sluggishness in binaural detection," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1709 – 1719, Apr. 1990.

[33] V. Best, F. J. Gallun, S. Carlile, and B. G. Shinn-Cunningham, "Binaural interference and auditory grouping," *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 1070–1076, Feb. 2007.

[34] T. Hirvonen and V. Pulkki, "Perceived distribution of horizontal ensemble of independent noise signals as function of sample length," in *Proc. of the 124th Intl. Conv. of the Audio. Eng. Soc.*, Amsterdam, the Netherlands, May 17 - 20 2008, Paper No. 7408.

[35] O. Santala and V. Pulkki, "Directional perception of distributed sound sources," *J. Acoust. Soc. Am.*, vol. 129, pp. 1522 – 1530, 2011.

[36] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975 – 979, 1953.

[37] A. S. Bregman, *Auditory Scene Analysis*, pp. 529 – 589, MIT Press, Cambridge, MA, 1st edition, 1994.

[38] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Am.*, vol. 105, no. 6, pp. 3436 – 3448, Jun. 1999.

[39] H. A. Witkin, S. Wapner, and T. Leventhal, "Sound localization with conflicting visual and auditory cues," *J. Exp. Psychol.*, vol. 43, no. 1, pp. 58 – 67, Jan. 1952.

[40] C. V. Jackson, "Visual factors in auditory localization," *Q. J. Exp. Psychol*, vol. 5, no. 2, pp. 52 – 65, 1953.

[41] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746 – 748, Dec. 1976.

[42] J. Merimaa and V. Pulkki, "Spatial impulse response rendering 1: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, December 2005.

[43] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[44] J. Vilkamo, T. Lokki, and V.Pulkki, "Directional audio coding: Virtual microphone based synthesis and subjective evaluation," *J. Audio Eng. Soc.*, vol. 57, no. 9, 2009.

[45] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, June 1997.

[46] M. Laitinen, T. Pihlajamäki, C. Erkut, and V. Pulkki, "Parametric time-frequency representation of spatial sound in virtual worlds," *ACM Transactions on Applied Perception (TAP)*, vol. 9, no. 2, pp. 8, 2012.

Sound localization in sagittal planes: Psychoacoustic foundation and models

Piotr Majdak, Robert Baumgartner, Bernhard Laback
Acoustics Research Institute, Austrian Academy of Sciences, Vienna

1 Sound localization in sagittal planes

1.1 Salient cues

Human normal-hearing (NH) listeners are able to localize sounds in space in terms of assigning direction and distance to the perceived auditory image [19]. Multiple mechanisms are used to estimate sound source direction in the three-dimensional space. While interaural differences in time and intensity are important for sound localization in the lateral dimension (left/right) [35], monaural spectral cues are assumed to be the most salient cues for sound localization in the sagittal planes (SPs) [20]. SPs are planes parallel to the median plane and include points of similar interaural time differences for a given distance. The *monaural* spectral cues are essential for the perception of the source elevation within a hemifield [17] and for front-back discrimination of the perceived auditory event [37].

Because interaural cues and monaural spectral cues are thought to be processed largely independently of each other [20], the interaural-polar coordinate system is often used to describe their respective contributions in the two dimensions. In the interaural-polar coordinate system the direction of a sound source is described with the lateral angle and the polar angle (see Fig. 1, left panel). SP localization refers to the listener's assignment of the polar angle for a given lateral angle and distance of the sound source.

Although spectral cues are processed monaurally, the information from both ears affects the perceived location in most cases [29]. The ipsilateral ear, namely, the one closer to the source, dominates and its relative contribution increases monotonically with increasing lateral angle [10]. If the lateral angle exceeds about 60°, the contribution of the

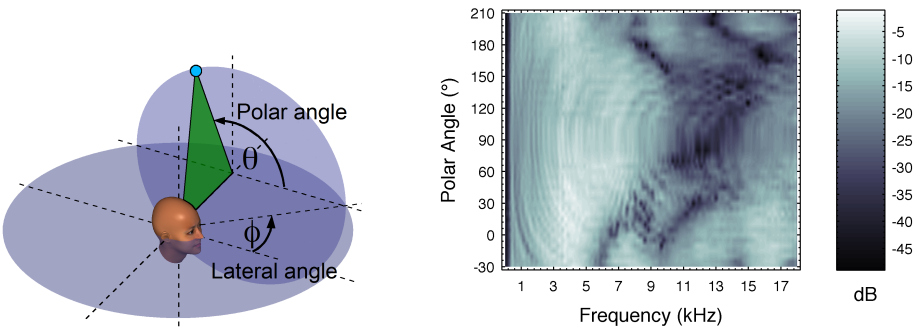

Figure 1: Left: Interaural-polar coordinate system. Right: HRTF magnitude spectra of a listener as a function of the polar angle in the median SP.

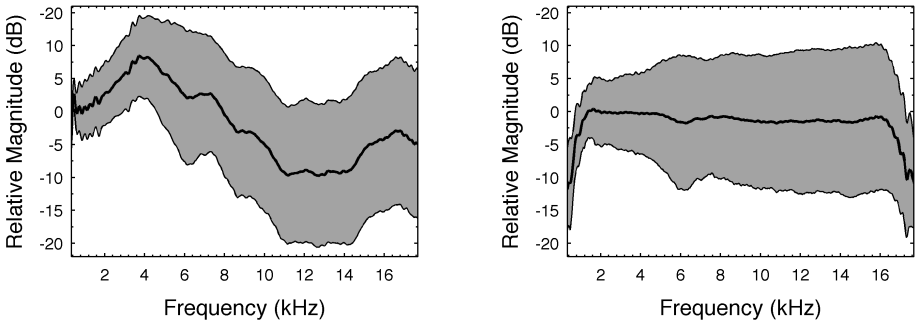

Figure 2: Left: Spatial variation of HRTFs around CTF for a listener. Right: Corresponding DTFs, that is, HRTFs with CTF removed. Solid line: Mean. Grey area: ± 1 std. dev.

contralateral ear becomes negligible. Thus, even for localization in the SPs, the lateral source position, mostly depending on the broadband binaural cues [20], must be known in order to determine the binaural weighting of the monaural cues.

The nature of the spectral features that are important for sound localization is still subject of investigations. Due to the physical dimensions, the pinna plays a larger role for the higher frequencies [26] and the torso for the lower frequencies [1]. Some psychoacoustic studies postulated that macroscopic patterns of the spectral features are important rather than fine spectral details [16, 17, 31, 13, 9]. On the other hand, other studies postulated that SP sound localization is possibly mediated by means of only a few local spectral features [27, 14, 37]. Despite a common agreement that the amount of the spectral features can be reduced without substantial reduction of the localization performance, the perceptual relevance of particular features has not been fully clarified yet.

1.2 Head-related transfer functions

The effect of the acoustic filtering of torso, head, and pinna can be described in terms of a linear time-invariant system by the so-called head-related transfer functions (HRTFs) [4, 32, 28]. HRTFs depend on the individual geometry of the listener and thus listener-specific HRTFs are required to achieve accurate localization performance for binaural synthesis [5, 25]. Usually, HRTFs are measured in an anechoic chamber by determining the acoustic response characteristics between loudspeakers at various directions and microphones inserted into the ear canals. Measured HRTFs contain both direction-dependent and direction-independent features and can be thought of as a series of two acoustic filters. The direction-independent filter, represented by the common transfer function (CTF), can be calculated from an HRTF set comprising many directions [24] by averaging the log-amplitude spectra of all available HRTFs of a listener's ear. The phase spectrum of the CTF is the minimum phase corresponding to the amplitude spectrum of the CTF.

Directional features are represented by the directional transfer functions (DTFs). The DTF for a particular direction is calculated by filtering the corresponding HRTF with the inverse CTF. The CTF usually exhibits a low-pass filter characteristic because the higher frequencies are attenuated for many directions due to the head and pinna shadow (see Fig. 2, left panel). Compared to HRTFs, DTFs usually pronounce frequencies and thus spectral features above 4 kHz (see Fig. 2, right panel). DTFs are commonly used to investigate the nature of spectral cues in SP localization experiments with virtual sources [24, 9, 22].

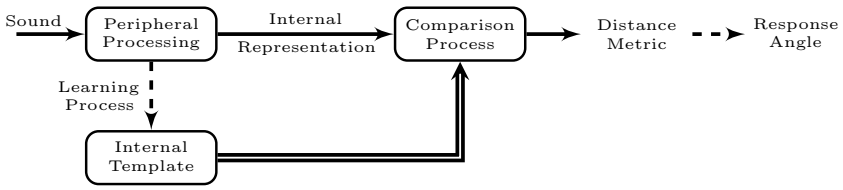

Figure 3: General structure of a template-based comparison model

2 Models of sagittal-plane localization

We consider models aiming at predicting the listener's polar response angle to the incoming sound [2]. Such models can help to explain psychoacoustic phenomena or to assess the spatial quality of audio systems while avoiding the running of costly and time-consuming localization experiments. We focus on a *functional* model where model parameters should correspond to physiological and/or psychophysical localization parameters. Until now, a functional model considering both spectral and temporal modulations exists only as a general concept [33]. Note that in order to address a particular research question, models dealing with specific types of modulations have been designed. For example, models for narrow-band sounds [27] were provided in order to explain the well-known effect of directional bands [4]. In order to achieve a sufficiently good prediction as an effect of the modification of the spectral cues, in [2], the incoming sound is assumed to be a *stationary broadband* signal, explicitly disregarding spectral and temporal modulations.

In general, machine learning approaches can be used to predict localization performance. Artificial neuronal networks (ANNs) have been shown to achieve rather accurate predictions when trained with large datasets of a single listener [15]. However, predictions for a larger subpopulation of human listeners would have required much more effort. Also, the interpretation of the ANN parameters is not straight forward. It is difficult to generalize the findings obtained with an ANN-based model to other signals, persons, and conditions and thus to better understand the mechanisms underlying spatial hearing.

Other localization models driven by various signal-processing approaches have been developed [3, 23]. These models are based on general principles of biological auditory systems, they do not, however, attempt to predict human-listener performance – their outcome shows rather the potential of the signal-processing algorithms involved.

2.1 Template-based comparison

A common property of existing sound localization models based on spectral cues is that they compare an internal representation of the incoming sound with a template [36, 11, 17] (see Fig. 3). The internal template is assumed to be created by means of learning the correspondence between the spectral features and direction of an acoustic event [12], based on feedback from other modalities. The localization performance is predicted by assuming that in the sound localization task, the comparison yields a distance metric that corresponds to the polar response angle of the listener. Thus, template-based models include a stage modeling the peripheral processing of the auditory system applied to both the template and incoming sound, and a stage modeling the comparison process in the brain.

Peripheral processing The peripheral processing stage is aimed at modeling the effect of human physiology while focusing on directional cues. The effect of the torso, head, and

outer ear are considered by filtering the incoming sound by an HRTF or a DTF. The effect of the ear canal, middle ear, and cochlear filtering can be considered by various model approximations. In the early HRTF-based localization models, a parabolic-shaped filter bank was applied [36]. Later, a filter bank averaging magnitude bins of the discrete Fourier transform of the incoming sound was used [17]. Both filter banks, while being computationally efficient, were drastically simplifying the auditory peripheral processing. The Gammatone (GT) filter bank [30] is a more physiology-related linear model of auditory filters and has been used in localization models [11]. A model accounting for the nonlinear effect of the cochlear compression is the dual-resonance nonlinear (DRNL) filter bank [18]. A DRNL filter consists of both a linear and a non-linear processing chain and is implemented by cascading GT filters and Butterworth low-pass filters, respectively. Another non-linear model uses a single main processing chain and accounts for the time-varying effects of the medial oliviocochlear reflex [38]. All those models account for the contribution of outer hair cells to a different degree and can be used to model the movements of the basilar membrane at a particular frequency.

The filter bank produces a signal for each center frequency and only the relevant frequency bands are considered in the model. Existing models used frequency bands with constant relative bandwidth on a logarithmic frequency scale [36, 17]. In model proposed in [2], the frequency spacing of the bands corresponds to one equivalent rectangular bandwidth (ERB) [8]. The lowest frequency is 0.7 kHz, corresponding to the minimum frequency thought to be affected by torso reflections [1]. The highest frequency considered in the model depends on the bandwidth of the incoming sound and is maximally 18 kHz, approximating the upper frequency limit of human hearing.

Further in the auditory system, the movements of the basilar membrane at each frequency band are translated into neural spikes by the inner hair cells (IHCs). An accurate IHC model has not been considered yet and does not seem to be vital for SP localization. Thus, different researches used different approximations. In our model, the IHC is modeled as half-wave rectification followed by a second order Butterworth low-pass with a cut-off frequency of 1 kHz [7]. Since the temporal effects of SP localization are not considered in [2], the output of each band is simply temporally averaged in terms of root mean square (RMS) amplitude, resulting in the internal representation of the sound. The same internal representation, and thus peripheral processing, is assumed for the template.

Comparison stage In the comparison stage, the internal representation of the incoming sound is compared with the internal template. Each entry of the template is selected by a polar angle denoted as template angle. A distance metric is calculated as a function of the template angle and can be interpreted as a potential descriptor for the response of the listener.

An early modeling approach proposed to compare the spectral derivatives of various orders in terms of a band-wise subtraction of the derivatives and then averaging over the bands [36]. The comparison of the first-order derivative corresponds to the assumption that the overall sound intensity does not contribute to the localization process. In the comparison of the second order derivatives, the differences in spectral tilt between the sound and the template do not contribute. Note that the plausibility of these comparison methods had not been investigated at that time. As another approach, Pearson's correlation has been proposed to evaluate the similarity between the sound and the template [27, 11]. Later, the inter-spectral differences (ISDs), namely, the differences between the internal representations of the incoming sound and the template calculated for each template angle and frequency band, were used [24] to show a correspondence between

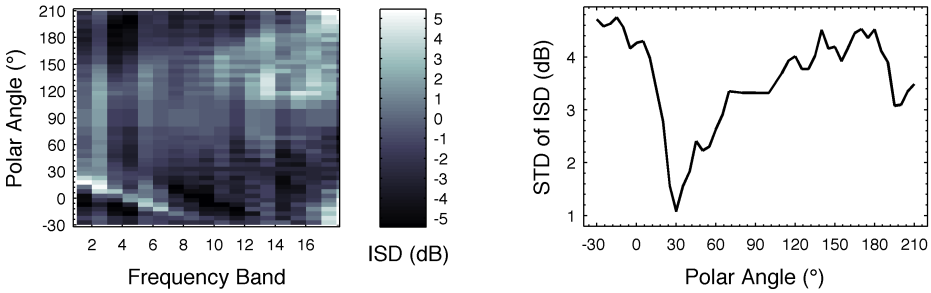

Figure 4: Example of the comparison process for a target polar angle of 30°. Left: ISDs as a function of the template angle. Right: Spectral standard deviation (STD) of ISDs as a function of the template angle.

the template angle yielding smallest spectral variance and the actual response of human listeners. All these comparison approaches were tested in [17], who, also distinguishing zeroth-, first-, and second-order derivatives of the internal representations, found that the standard deviation of ISDs best described their results. This configuration corresponds to an average of the first-order derivative from [36], which is robust against changes in the overall level in the comparison process.

In [2], ISDs are calculated for a template angle and for each frequency band (see Fig. 4, left panel). Then, the spectral standard deviations of ISDs are calculated for all available template angles (see Fig. 4, right panel). For band-limited sounds, the internal representation results in an abrupt change at the cut-off frequency of the sound. This change affects the standard deviation of the ISDs. Thus, in [2], the ISDs are calculated only within the bandwidth of the incoming sound.

The result of the comparison stage is a distance metric corresponding to the prediction of the polar response angle. Early modeling approaches used the minimum distance to determine the predicted response angle [36], which would nicely fit the minimum of the distance metric used in our example (see Fig. 4, right panel). Also, the cross-correlation coefficient has been used as a distance metric and its maximum has been interpreted as the prediction of the response angle [27]. Both approaches represent a deterministic interpretation of the distance metric, resulting in exactly the same predictions for the same sounds. This is rather unrealistic. Subjects, repeatedly listening to the same sound, often do not respond to exactly the same direction [6]. The actual responses are known to be scattered and can be even multimodal. The scatter of one mode can be described by the Kent distribution [6], which is an elliptical probability distribution on the two-dimensional unit sphere.

2.2 Response probability

In order to model the probabilistic response pattern of listeners, a mapping of the distance metric to polar-response probabilities via similarity indices (SIs) has been proposed [17]. For a particular target angle and ear, a monaural SI has been obtained by using the distance metric as the argument of a Gaussian function with a mean of zero and a standard deviation of two (Fig. 5, $U = 2$). While this choice appears to be somewhat arbitrary, it models the probabilistic relation between the distance metric and the probability of responding to a given direction. Note that the resulting SI is bounded by zero and one and valid for the analysis of the incoming sound at one ear only.

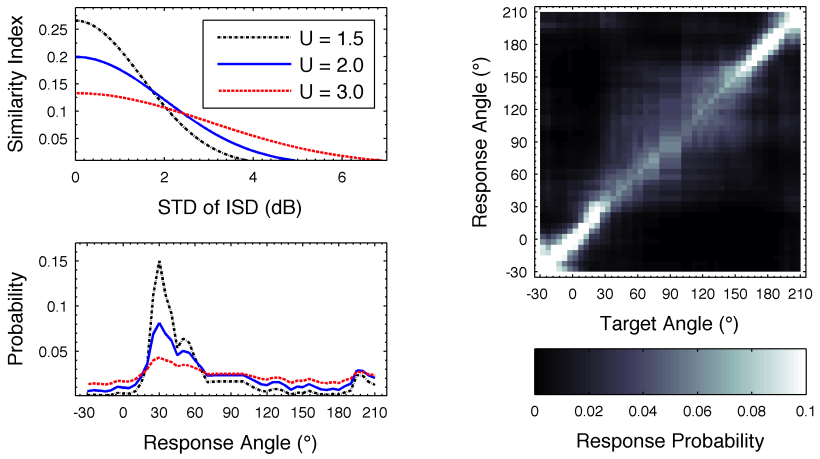

Figure 5: Left column: Mapping function of SI (top panel) for various uncertainties U and the resulting PMVs (bottom panel) corresponding to the example shown in Fig. 4. Right column: Prediction matrix; predicted response PMV of the localization model as a function of the target angle for the baseline condition in the median SP.

The width of the mapping function, U in Fig. 5, actually reflects a property of an individual listener. A listener being more precise in the response to the same sound would need a more narrow mapping than a less precise listener. Thus, in contrast to the previous approach [17], in our model we consider the width of the mapping function as a listener-specific uncertainty, U. It accounts for listener-specific localization precision [24, 37] due to reasons like training and attention [12]. Note that for simplicity, direction-dependent response precision is neglected. The lower the uncertainty, U, the higher the assumed sensitivity of the listener to distinguish spectral features. In [2], this parameter is used to calibrate the model to listener-specific performance.

The model stages described so far are monaural. Thus, they do not consider binaural cues and have been designed for the median SP where the interaural differences are zero and thus binaural cues do not contribute. In order to take into account the contribution of both ears, the monaural model results for both ears are combined. Previous approaches averaged the monaural SIs for both ears [17] and thus were able to consider the contribution of both ears for targets placed in the median SP. In our model, we extend the lateral target range to arbitrary SPs by applying a binaural weighting function [10, 21], which reduces the contribution of the contralateral ear depending on the perceived lateral direction of the target sound. Thus, the binaural weighting function is applied to each monaural SI, and the sum of the weighted monaural SIs yields the binaural SI.

For an incoming sound, the binaural SIs are calculated for all template entries selected by the template angle. Such a binaural SI as a function of the template angle is related to the listener's response probability as a function of the response angle. It can be interpreted as a discrete version of a probability density function, namely, a probability mass vector (PMV), showing the probability of responding at an angle to a particular target. In order to obtain a PMV, the binaural SI is normalized to have a sum of one.

The PMVs, calculated separately for each target under consideration, are represented in a prediction matrix. This matrix describes the probability of responding at a polar angle given a target placed at a specific angle. The right panel of Fig. 5 shows the prediction matrix resulting for the exemplary listener in a baseline condition where the listener uses his/her own DTFs, and all available listener-specific DTFs are used as targets. The abscissa shows the target angle, the ordinate shows the response angle, and the brightness

represents the response probability. This representation allows for a visual comparison between the model predictions and the responses obtained from actual localization experiments.

2.3 Interpretation of the probabilistic model predictions

In order to compare the probabilistic results from the model with the experimental results, likelihood statistics, calculated for actual responses from sound localization experiments and for responses resulting from virtual experiments driven by the model prediction, can be used [Eq. (1) in 17]. The comparison between the two likelihoods allows one to evaluate the validity of the model, because only for similar likelihoods the model is assumed to yield valid predictions. The likelihood has, however, a weak correspondence with localization performance parameters commonly used in psychophysics.

Localization performance in the polar dimension usually considers local errors and hemifield confusions [25]. Although these errors derived by geometrical aspects cannot sufficiently represent our current understanding of human hearing, they are frequently used and thus enable comparison of results between studies. Quadrant errors (QE), that is the percentage of polar errors larger or equal to 90°, represent the confusions between hemifields (e.g., front/back or up/down) without considering the local response pattern. Unimodal local responses can be represented as a Kent distribution [6], which, considering the polar dimension only, can be approximated by the polar bias and polar variance. Thus, the local errors are calculated only for local responses within the correct hemifield, namely, without the responses yielding the QEs. A single representation of the local errors is the local polar RMS error (PE), which combines localization bias and variance in a single metric.

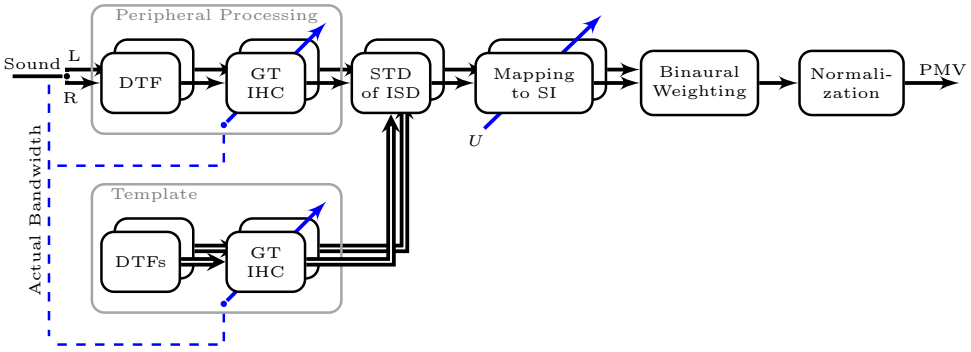

Figure 6: Structure of the SP localization model from [2].

In [2], QEs and PEs are calculated from the PMVs. The QE is the sum of the PMV entries outside the local polar range for which the response-target difference is greater or equal to 90°. The PE is the discrete expectancy value within the local polar range. In the visualization of prediction matrices (see for example right column of Fig. 5), bright areas in the upper left and bottom right corners would indicate large QEs, a strong concentration of the brightness at the diagonal would indicate small PEs. Both errors can be calculated either for a specific target angle or as the arithmetic average across all target angles considered in the prediction matrix.

Figure 6 summarizes the final structure of the model proposed in [2]. It requires the incoming signal from a sound source as the input and results in the response probability as a function of response angle (PMV) for given template DTFs. Then, from PMVs calculated for the available target angles, QEs and PEs are calculated for a direct comparison

with the outcome of a sound localization experiment. Figure 7 shows the quality of the predictions in terms of correlation between the actual localization performance obtained in sound localization experiments and the modeled localization performance.

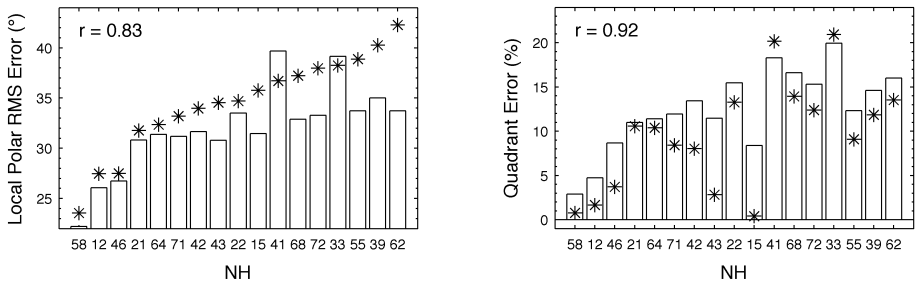

Figure 7: Localization performance (PE, QE) predicted by the model for listener-specific calibration (bars) and actual performance obtained in sound localization experiments (asterisks). r: Pearson's correlation coefficient with respect to actual and predicted performance.

3 Summary

Sound localization in sagittal planes refers to the ability to estimate the sound-source elevation and to distinguish between front and back. It relies on monaural spectral features encoded in the HRTFs. The SP localization performance is usually measured in time-consuming experiments. The model [2] can be applied to predict the SP localization performance of individual listeners. It is based on a template-based comparison and uses a listener-specific calibration. The potential applications are, among others:

1. The evaluation of the spatial quality of binaural recordings
2. The assessment of the spatial quality of directional cues provided by the microphone placement in hearing-assist devices
3. The evaluation of the sagittal-plane localization in surround-sound systems

In order to further demonstrate the model, predictions for exemplary applications are shown in the following figures.

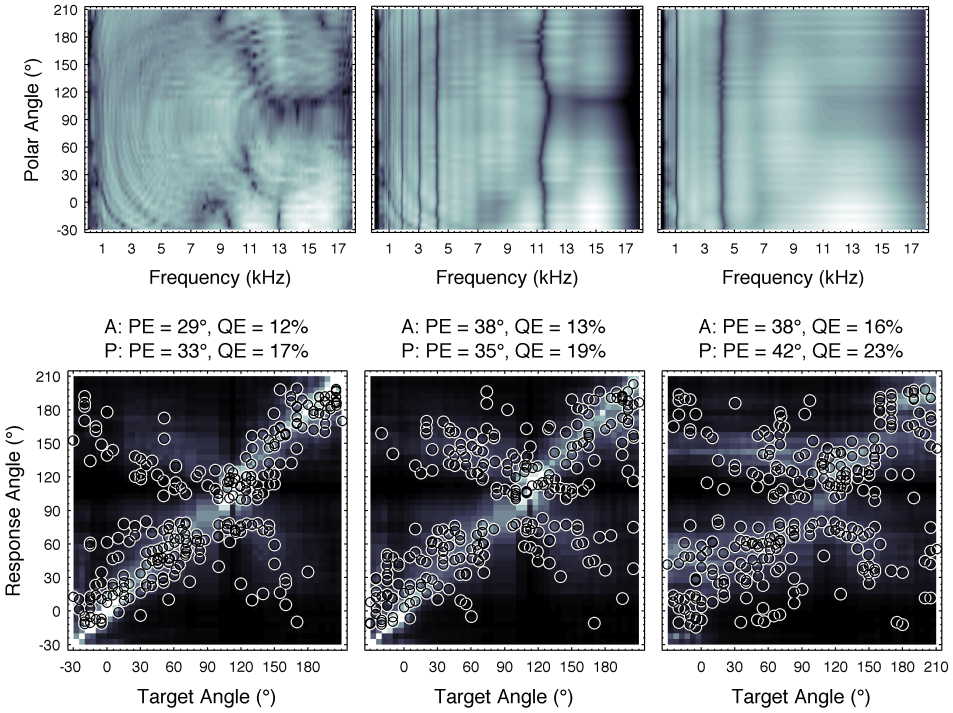

Figure 8: Effect of the number of spectral channels. Top row: Channelized DTFs of median SP from [9]. Bottom row: Prediction matrices and actual responses (open circles). Left column: Unlimited number of channels. Center column: 24 spectral channels. Right column: 9 spectral channels. A: Actual performance from [9]. P: Predicted performance.

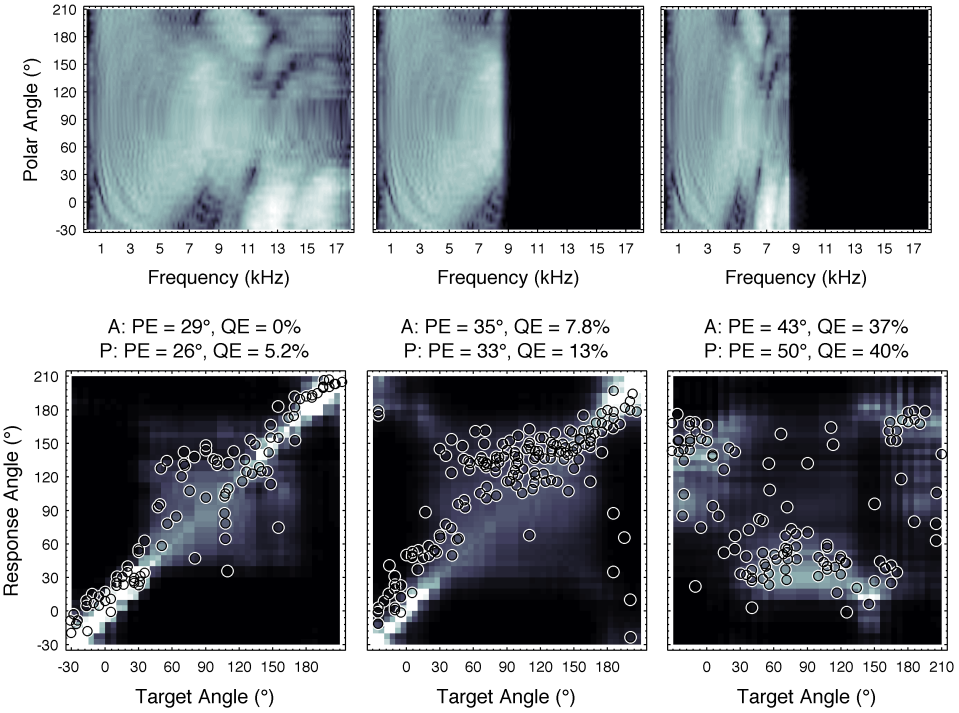

Figure 9: Effect of drastic HRTF modifications. Localization with the original (left column), band-limited (center column), and spectrally warped (right column) DTFs. Top row: DTFs in the median SP. Bottom row: Prediction matrices. Open circles: actual responses from [34].

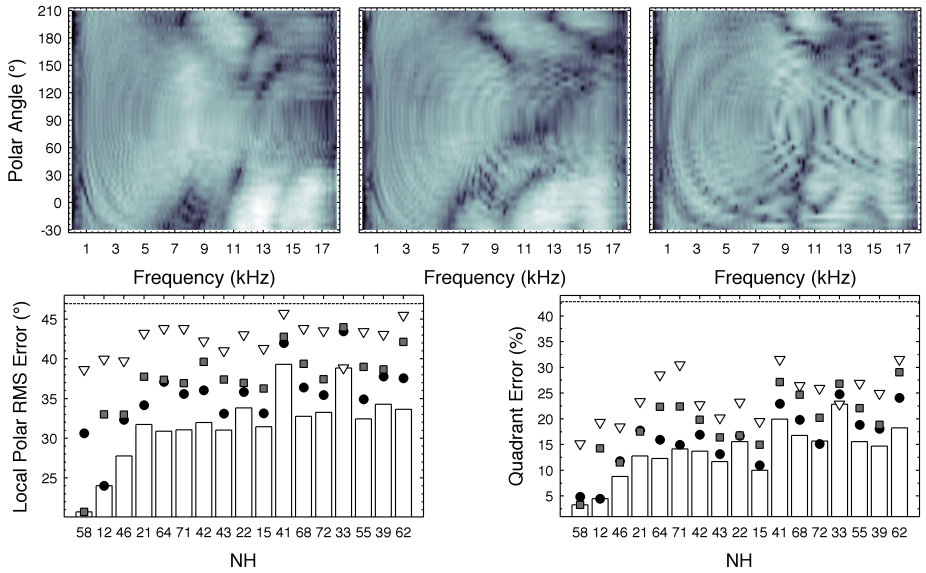

Figure 10: Effect of listening with others ears. Top panels: DTFs of median SP; NH12 (left), NH58 (center), NH33 (right). Brightness: Spectral magnitude. Bottom panels: Localization performance (bars) of the pool listening to DTFs of NH12 (circles), NH58 (squares), NH33 (triangles). Dashed line: Chance performance.

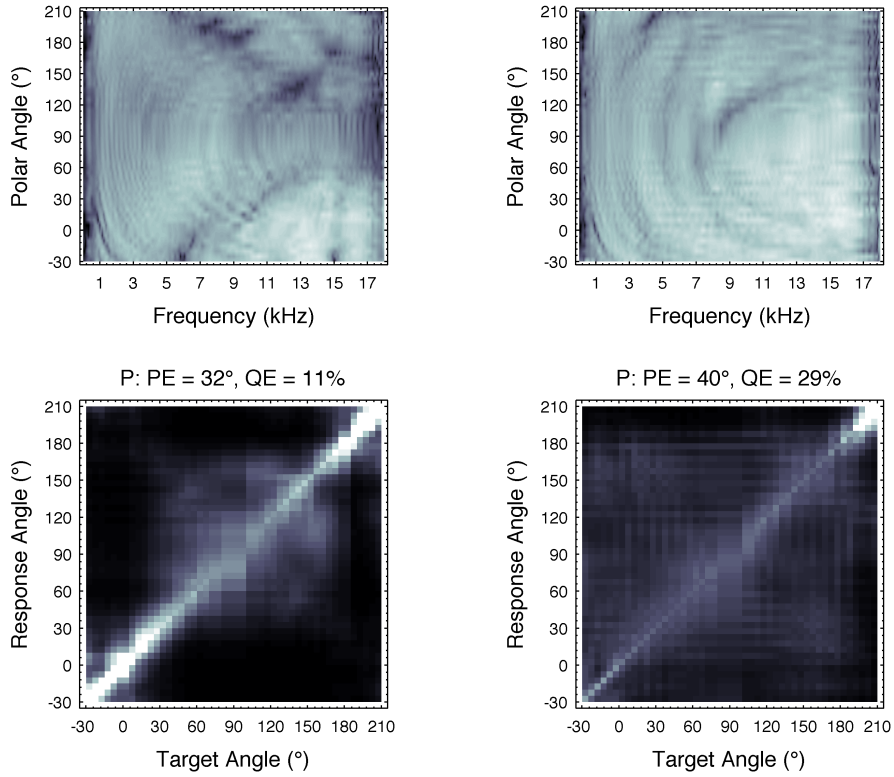

Figure 11: Effect of the microphone placement. Left column: In-the-ear microphone. Right column: Behind-the-ear microphone.

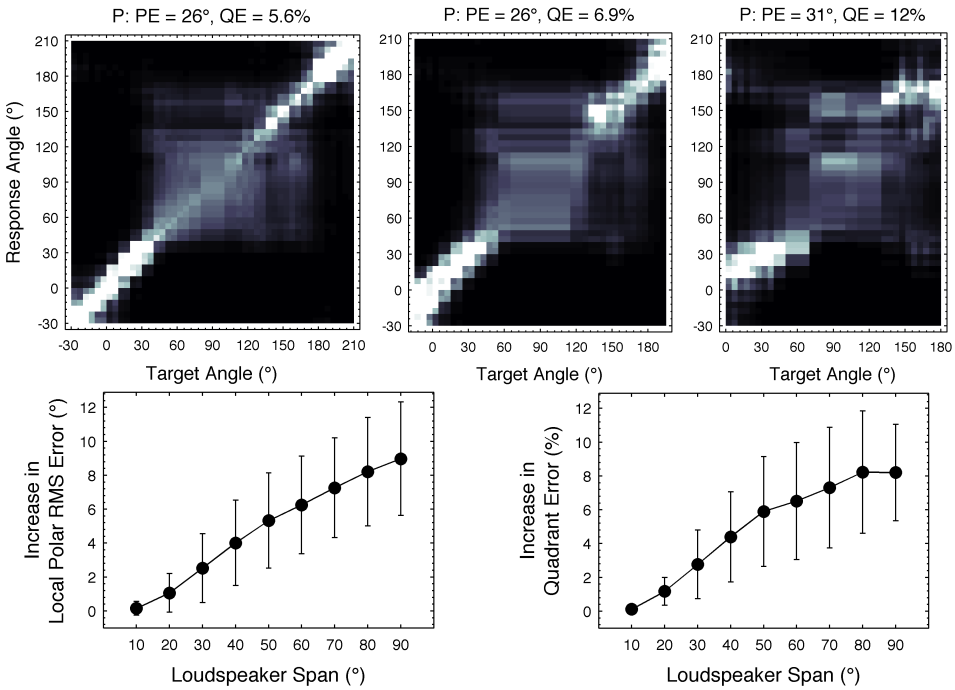

Figure 12: Vector-base amplitude panning for different loudspeaker spans. Left: Span of 0° (single-loudspeaker synthesis, baseline condition). Center: Span of 30°. Right: Span of 60°. Bottom panels: Increase in localization errors as a function of the loudspeaker span.

References

[1] Algazi, V. R., Avendano, C., and Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. *J Acoust Soc Am*, 109:1110–1122.

[2] Baumgartner, R., Majdak, P., and Laback, B. (2013). Assessment of sagittal-plane sound-localization performance in spatial-audio applications. In Blauert, J., editor, *The technology of binaural listening*, chapter 4. Springer, Berlin–Heidelberg–New York NY.

[3] Blanco-Martin, E., Casajus-Quiros, F. J., Gomez-Alfageme, J. J., and Ortiz-Berenguer, L. I. (2010). Estimation of the direction of auditory events in the median plane. *Appl Acoust*, 71:1211–1216.

[4] Blauert, J. (1974). *Räumliches Hören (Spatial Hearing)*. S. Hirzel Verlag Stuttgart.

[5] Bronkhorst, A. W. (1995). Localization of real and virtual sound sources. *J Acoust Soc Am*, 98:2542–2553.

[6] Carlile, S., Leong, P., and Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. *Hear Res*, 114:179–196.

[7] Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J Acoust Soc Am*, 99:3615–3622.

[8] Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes form notched-noise data. *Hear Res*, 47:103–138.

[9] Goupell, M. J., Majdak, P., and Laback, B. (2010). Median-plane sound localization as a function of the number of spectral channels using a channel vocoder. *J Acoust Soc Am*, 127:990–1001.

[10] Hofman, M. and Van Opstal, J. (2003). Binaural weighting of pinna cues in human sound localization. *Exp Brain Res*, 148:458–470.

[11] Hofman, P. M. and Opstal, A. J. V. (1998). Spectro-temporal factors in two-dimensional human sound localization. *J Acoust Soc Am*, 103:2634–2648.

[12] Hofman, P. M., van Riswick, J. G. A., and van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neurosci*, 1:417–421.

[13] Hwang, S. and Park, Y. (2008). Interpretations on pricipal components analysis of head-related impulse responses in the median plane. *J Acoust Soc Am*, 123:EL65–EL71.

[14] Iida, K., Itoh, M., Itagaki, A., and Morimoto, M. (2007). Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Appl Acoust*, 68:835–850.

[15] Jin, C., Schenkel, M., and Carlile, S. (2000). Neural system identification model of human sound localization. *J Acoust Soc Am*, 108:1215–1235.

[16] Kulkarni, A. and Colburn, H. S. (1998). Role of spectral detail in sound-source localization. *Nature*, 396:747–749.

[17] Langendijk, E. H. A. and Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *J Acoust Soc Am*, 112:1583–1596.

[18] Lopez-Poveda, E. A. and Meddis, R. (2001). A human nonlinear cochlear filterbank. *J Acoust Soc Am*, 110:3107–3118.

[19] Lord Rayleigh, F. R. S. (1907). On our perception of sound direction. *Philos Mag*, 13:214–232.

[20] Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *J Acoust Soc Am*, 111:2219–2236.

[21] Macpherson, E. A. and Sabin, A. T. (2007). Binaural weighting of monaural spectral cues for sound localization. *J Acoust Soc Am*, 121:3677–3688.

[22] Majdak, P., Goupell, M. J., and Laback, B. (2010). 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Atten Percept Psycho*, 72:454–469.

[23] May, T., van de Par, S., and Kohlrausch, A. (2011). A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans Audio Speech Lang Proc*, 19:1–13.

[24] Middlebrooks, J. C. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. *J Acoust Soc Am*, 106:1480–1492.

[25] Middlebrooks, J. C. (1999b). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *J Acoust Soc Am*, 106:1493–1510.

[26] Middlebrooks, J. C. and Green, D. M. (1991). Sound localization by human listeners. *Annu Rev Psychol*, 42:135–159.

[27] Middlebrooks, J. C. and Green, D. M. (1992). Observations on a principal components analysis of head-related transfer functions. *J Acoust Soc Am*, 92:597–599.

[28] Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *J Audio Eng Soc*, 43:300–321.

[29] Morimoto, M. (2001). The contribution of two ears to the perception of vertical angle in sagittal planes. *J Acoust Soc Am*, 109:1596–1603.

[30] Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). *An efficient auditory filterbank based on the gammatone function.* APU, Cambridge, UK.

[31] Senova, M. A., McAnally, K. I., and Martin, R. L. (2002). Localization of virtual sound as a function of head-related impulse response duration. *J Audio Eng Soc*, 50:57–66.

[32] Shaw, E. A. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *J Acoust Soc Am*, 56:1848–1861.

[33] Vliegen, J. and Opstal, A. J. V. (2004). The influence of duration and level on human sound localization. *J Acoust Soc Am*, 115:1705–1703.

[34] Walder, T. (2010). Schallquellenlokalisation mittels Frequenzbereich-Kompression der Außenohrübertragungsfunktionen (sound-source localization through warped head-related transfer functions). Master's thesis, University of Music and Performing Arts, Graz, Austria.

[35] Wightman, F. L. and Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *J Acoust Soc Am*, 91:1648–1661.

[36] Zakarauskas, P. and Cynader, M. S. (1993). A computational theory of spectral cue localization. *J Acoust Soc Am*, 94:1323–1331.

[37] Zhang, P. X. and Hartmann, W. M. (2010). On the ability of human listeners to distinguish between front and back. *Hear Res*, 260:30–46.

[38] Zilany, M. S. A. and Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J Acoust Soc Am*, 120:1446–1466.

SUPER-RESOLUTION SOUND FIELD ANALYSES

C.T. Jin and N. Epain

CARLab, School of Electrical and Information Engineering
The University of Sydney
Sydney, Australia

1 Introduction

Spherical Microphone Arrays (SMAs) have become increasingly popular during the past decade. One of the unique advantages of SMAs is they provide a uniform panoramic view of the sound field. This capability motivates their use for the spatial analysis of sound fields. In [1] and [2], for instance, standard beamforming methods are employed to provide an energy map of the sound field recorded by an SMA. Recently, more elaborate techniques such as EB-MUSIC [3] or EB-ESPRIT [4] (spherical harmonic domain implementations of MUSIC [5] and ESPRIT [6]) have been applied to sound field imaging.

The issue with the above sound field analysis methods is that they are bounded by the SMA's intrinsic resolution. In practice, SMAs are typically comprised of a few dozen microphones. Thus the sound field images obtained from SMA recordings have limited resolution. In recent work [7, 8, 9], techniques based on sparse recovery (SR) have been proposed that enable an increase in the spatial resolution of the sound field description. When these techniques are applied to sound scene analysis or for playing back and listening to HOA sound scenes, we refer to the analysis as super-resolution imaging or sound scene upscaling, respectively. These techniques are based on the hypothesis that the sound field can be approximated by a few dominant plane-wave sources. Although this is a fair assumption in simple free-field scenarios, it is clearly not true in the presence of noise or reverberation. Nevertheless, this tutorial shows how to effectively deal with non-sparse sound conditions using subspace pre-processing.

In summary, this tutorial briefly reviews sound field analysis in the spherical or HOA domain. It then proceeds to describe the sparse recovery (SR) problem that leads to super-resolution imaging and presents an effective pre-processing method that improves the robustness of SR-based methods in the presence of noise or reverberation.

1.1 Spherical harmonic expansion of a sound field

In the frequency domain, any sound field consisting of incident sound waves can be expressed as the following series [10]:

$$p\left(r,\vartheta,\varphi,f\right)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l} i^{l} j_{l}(kr)\, Y_{l}^{m}(\vartheta,\varphi)\, b_{lm}(f)\,, \tag{1}$$

where $p\left(r,\vartheta,\varphi,f\right)$ is the acoustic pressure for frequency f at the point in space with spherical coordinates (r,ϑ,φ), i is the imaginary unit, j_l is the spherical Bessel function of degree l, Y_l^m is the spherical harmonic function of order l and degree m, $b_{lm}(f)$ is the spherical harmonic expansion coefficient for order l, degree m, frequency f, k is the wave number, $k=2\pi f/c$, and c is the speed of sound. We refer to Eq. (1) as the Bessel-weighted spherical harmonic expansion of the sound field.

1.2 Higher Order Ambisonics

Truncating the series in Eq. (1) to order L results in a good approximation of the acoustic pressure within a sphere of radius $\hat{r}$ centered on the origin [11]:

$$p\left(r\leq\hat{r},\vartheta,\varphi,f\right)\approx\sum_{l=0}^{L}\sum_{m=-l}^{l} i^{l} j_{l}(kr)\, Y_{l}^{m}(\vartheta,\varphi)\, b_{lm}(f) \quad \text{with } \hat{r}=\frac{2L+1}{ek}\,, \tag{2}$$

where e is the mathematical constant known as Euler's number.

As shown in Eq. (2), the coefficients $b_{lm}(f)$ of the Bessel-weighted spherical harmonic transform suffice to describe the sound field around the origin. This transformation underlies the Higher Order Ambisonics (HOA) sound field reproduction method [12] and the coefficients $b_{lm}(f)$ are referred to as the frequency domain HOA signals. The order-L HOA time domain signals are obtained from the $b_{lm}(f)$ coefficients using an inverse Fourier transform and provide a compact and scalable description of the sound field. HOA signals can be decoded for playing back over various loudspeaker configurations or over headphones.

1.3 Recording HOA Signals with Spherical Microphone Arrays

An advantageous feature of spherical microphone arrays leading to their recent popularity is their ability to record HOA signals. Obtaining HOA signals from the microphone signals is referred to as *HOA-encoding*. We briefly describe the encoding process and leave the details for the appendix. In the following, we assume the microphone array has P microphones. In practice, the time-domain, order-L HOA signals are obtained from the microphone signals using a matrix of finite impulse response (FIR) encoding filters, $\mathbf{E}(t)$. The time-domain HOA encoding operation is expressed by the formula:

$$\hat{\mathbf{b}}(t) = \mathbf{E}(t) \circledast \mathbf{x}(t) \,, \tag{3}$$

where $\hat{\mathbf{b}}(t)$ denotes the vector of the time-domain *encoded* HOA signals, $\mathbf{x}(t)$ denotes the vector of the time-domain microphone signals, and $\circledast$ denotes the convolution of a vector of signals by a matrix of filters, *i.e.*:

$$\hat{b}_{lm}(t) = \sum_{p=1}^{P} e_{lm,p}(t) * x_p(t) \,, \tag{4}$$

where $e_{lm,p}(t)$ is the encoding filter corresponding to the HOA signal of order l and degree m and the p-th microphone, and $*$ is the convolution product. The derivation of the HOA encoding filters is presented in the appendix for completeness, but is not significant for the purposes of this tutorial.

2 Sound field imaging with Spherical Microphone Arrays

In the SMA framework, we have seen that the signals recorded by the microphones can be transformed into a finite-order spherical harmonic representation of the sound field, the resolution of which depends on the order: the higher the order, the finer the resolution. Practical considerations generally limit the order of current SMAs. In any case, for this tutorial our starting point shall be the matrix, $\mathbf{B}_L$, of the order-L HOA signals with sample length N:

$$\mathbf{B}_L = [\mathbf{b}_{0,0}, \mathbf{b}_{1,-1}, ..., \mathbf{b}_{L,L}]^\mathsf{T} \,, \quad \text{where} \quad \mathbf{b}_{l,m} = [b_{l,m}(1), b_{l,m}(2), ..., b_{l,m}(N)]^\mathsf{T} \,, \tag{5}$$

and $b_{l,m}(n)$ denotes the n-th time sample of the order-l, degree-m HOA signal.

We now consider the task of deciphering: "what sounds come from where." This is the task of sound field imaging, it is fundamental to acoustic recordings and consists in representing the acoustic energy as a function of the incoming direction and frequency. The traditional approach to estimate the energy corresponding to a particular direction consists in steering a beam towards this direction and calculating the energy of the beamformer's output [1, 2]. In the HOA domain, beamforming is accomplished by multiplying a spherical harmonic steering vector, $\mathbf{y}_L(u)$, with the matrix of HOA signals. For example, we estimate the energy , $e_L(u)$, incoming from direction (θ_u, ϕ_u) as:

$$e_L(u) = \left\| \frac{1}{L+1} \, \mathbf{y}_L(u)^\mathsf{T} \, \mathbf{B}_L \right\|^2 \,, \tag{6}$$

where $\mathbf{y}_L(u)$ is the spherical harmonic steering vector of the up-to-order-L spherical harmonic function values in the direction (θ_u, ϕ_u):

$$\mathbf{y}_L(u) = \left[Y_0^0(\theta_u, \phi_u), Y_1^{-1}(\theta_u, \phi_u), Y_1^0(\theta_u, \phi_u), ..., Y_L^L(\theta_u, \phi_u) \right]^\mathsf{T} \,, \tag{7}$$

and Y_l^m denotes the order-l, degree-m real-valued spherical harmonic function.

As mentioned earlier, practical and physical constraints generally limit the order of the HOA signals recorded by SMAs (typically up to order 3 or 4) and this in turn limits their resolution. In the next section, we show how considerations that impose sparsity on the sources generating the sound field improves the resolution of the acoustic imaging.

3 Super-resolution imaging

The super-resolution imaging technique presented here is closely related to the compressed sensing literature and employs sparse recovery (SR) techniques [7, 8, 9]. In the SR approach, we consider a plane-wave decomposition of the HOA signals. In other words a matrix of plane-wave signals, $\mathbf{X}$, is calculated such that:

$$\mathbf{D}_L \mathbf{X} = \mathbf{B}_L \,, \tag{8}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_U]^\mathsf{T}$, $\mathbf{D}_L$ is the matrix expressing the contribution of each plane-wave to the HOA signals:

$$\mathbf{D}_L = [\mathbf{y}_L(1), \mathbf{y}_L(2), ..., \mathbf{y}_L(U)] \,, \tag{9}$$

and U is the number of plane wave directions. We refer to $\mathbf{D}_L$ as the plane-wave dictionary, where the number of entries, U, determines the resolution. As the aim is to obtain a spatially sharp description of the sound field, the dictionary is chosen with a number of directions far greater than the number of HOA signals. In this tutorial, we use a dictionary of 2562 directions derived by repeated subdivisions of the faces of an icosahedron.

The classical method for solving for $\mathbf{X}$ in Equation (8) is to select the solution with the least-square norm. The problem with this solution is that it is also the one that distributes the sound energy the most evenly across directions, which is physically incompatible with the assumption of discrete sound sources. Thus, the least-square norm results in a blurred image. The SR approach, on the other hand, tries to select the solution with the fewest active plane waves, the *sparsest* solution. In practice, it is shown in the compressed sensing literature [13] that we obtain such a solution by solving the following optimization problem:

$$\text{minimize} \quad \|\mathbf{X}\|_{1,2} \quad \text{subject to} \quad \mathbf{D}_L\mathbf{X} = \mathbf{B}_L \,, \tag{10}$$

where $\|\cdot\|_{1,2}$ denotes the $L1,2$-norm, defined by:

$$\|\mathbf{X}\|_{1,2} = \sum_{u=1}^{U} \sqrt{\sum_{n=1}^{N} x_u(n)^2} \,. \tag{11}$$

The fact that this optimization problem provides a sparse solution is by no means obvious. The proof is beyond the scope of this tutorial and requires a study of the compressed sensing literature.

In this paper we solve Problem (10) using an Iteratively-Reweighted Least-Square (IRLS) algorithm [14]. Solving Problem (10) results in a sparse set of signals and their corresponding directions selected from the plane-wave dictionary. The sparse set of plane-wave signals can then be used to estimate the order-L' HOA signals with $L' > L$. We refer to this operation as *upscaling* the HOA signals. Upscaling the signals to order L' is achieved by simply multiplying the plane-wave signals by the order-L' dictionary, $\mathbf{D}_{L'}$, defined similarly as $\mathbf{D}_L$. Note that the plane-wave dictionary, $\mathbf{D}$, is easily upscaled because the columns are the spherical harmonic representation of plane-waves in a given direction and can be represented to any order mathematically. In other words, the matrix of the upscaled HOA signals, $\mathbf{B}_{L\rightarrow L'}$ is given by:

$$\mathbf{B}_{L\rightarrow L'} = \mathbf{D}_{L'}\mathbf{X} \,. \tag{12}$$

A *super-resolution* acoustic energy map can then be estimated from the upscaled HOA signals, using the standard beamforming technique described in Section 2 (Equation 6). Or if the objective is listening to the separated signals, super-resolution beamforming toward the target signal using the upscaled HOA signals improves the signal separation.

4 Super resolution imaging with subspace pre-processing

The SR approach presented in the previous section is based on the assumption that the sound field is sparse when expressed using a plane wave basis. However, recorded HOA signals are always polluted by measurement noise which is *not* sparse in the plane-wave domain. Further, the sound field may consist of a mixture of a few spatially discrete sources, together with a diffuse sound field comprising the reverberated sound waves.

The trick to make the super-resolution imaging more robust in the presence of diffuse sound or measurement noise is to separate the HOA signals into two parts, a "directional" and "diffuse" component, prior to the imaging. Our approach for extracting the diffuse component of the sound field is to project the HOA signals onto a subspace that is orthogonal (or mostly orthogonal) to the directional component of the sound field.

We first consider the correlation matrix of the order-L HOA signals, $\mathbf{C}_L$, which can be decomposed in terms of its eigenvalues and eigenvectors:

$$\mathbf{C}_L = \mathbf{B}_L\,\mathbf{B}_L^{\mathsf{T}} = \mathbf{V}\,\mathbf{S}\,\mathbf{V}^{\mathsf{T}} \,, \tag{13}$$

where $\mathbf{V}$ is the matrix of the eigenvectors and $\mathbf{S}$ is the diagonal matrix of the eigenvalues:

$$\mathbf{S} = \text{diag}\left(\left[s_1, s_2, s_3, ..., s_{(L+1)^2}\right]\right), \quad \text{where} \;\; s_1 \geq s_2 \geq s_3 \geq ... \geq s_{(L+1)^2} \,. \tag{14}$$

We define the "diffuse separation" matrix, $\mathbf{A}$, as:

$$\mathbf{A} = \mathbf{V}\,\mathbf{\Gamma}\,\mathbf{V}^{\mathsf{T}} \,. \tag{15}$$

The matrix $\mathbf{\Gamma}$ is a diagonal matrix of weights comprised between 0 and 1:

$$\mathbf{\Gamma} = \text{diag}\left(\left[\gamma_1, \gamma_2, \gamma_3, ..., \gamma_{(L+1)^2}\right]\right), \quad \text{where} \;\; \gamma_i = \sin\left(\frac{\pi s_{(L+1)^2}}{2 s_i}\right)^2 \,. \tag{16}$$

The effect of $\mathbf{A}$ operating on the HOA signals is to: 1) Project the HOA signals onto the eigenvectors; 2) Weight the obtained signals, such that the diffuse components of the sound field are preserved, while the direct components are discarded;

3) Return to the HOA domain. The idea here is to compare the various eigenvalues against the smallest eigenvalue, $s_{(L+1)^2}$, because the small eigenvalues correspond mostly to diffuse or noise components, whereas large eigenvalues correspond to the directional components. It is important to note here that, in contrast to the MUSIC algorithm [5], no assumption is made on the number of sound sources present in the directional component of the sound field.

We now describe the method for super-resolution imaging with subspace pre-processing. The first step is to extract the diffuse component, $\mathbf{B}_L^{(\mathrm{dif})}$, and directional component, $\mathbf{B}_L^{(\mathrm{dir})}$, of the HOA signals as shown below:

$$\mathbf{B}_L^{(\mathrm{dif})} = \mathbf{A}\,\mathbf{B}_L\,, \qquad \mathbf{B}_L^{(\mathrm{dir})} = (\mathbf{I}-\mathbf{A})\,\mathbf{B}_L\,. \tag{17}$$

The signals in $\mathbf{B}_L^{(\mathrm{dir})}$ contain less noise or diffuse sound than the original signals and therefore are sparser in the plane wave source domain. The directional and diffuse components are processed separately. The second step is to apply the SR plane-wave decomposition (Section 3) to the directional component to obtain the upscaled signals $\mathbf{B}_{L\to L'}^{(\mathrm{dir})}$. The third step is to form a composite energy map comprised of a super-resolution map for the directional component and a low-resolution map for the diffuse component. The energy of the composite map for direction (θ_u, ϕ_u), $\hat{e}_{L\to L'}(u)$, is given by:

$$\hat{e}_{L\to L'}(u) = e_{L\to L'}^{(\mathrm{dir})}(u) + e_L^{(\mathrm{dif})}(u)\,. \tag{18}$$

where $e_{L\to L'}^{(\mathrm{dir})}(u)$ and $e_L^{(\mathrm{dif})}(u)$ are the energies of the directional (order-L') and diffuse (order-L) maps, obtained as described in Section 2 (Equation 6). Note that the diffuse component can be excluded when it is expected to contain mostly measurement noise. However it may be useful to analyze this component in situations where reverberation or other diffuse sounds are present.

5 Numerical simulations

In this section we present numerical simulation results validating the imaging method described above. In the simulation, we consider a sound field consisting of 6 plane waves with a diffuse background. The plane waves are incoming from the directions shown in Table 1. The corresponding waveforms are uncorrelated Gaussian white noise signals with equal energy. The diffuse component of the sound field is composed by a very large number of plane-wave, Gaussian, white noise signals with equal energies, that are evenly distributed across space. As ground truth we use the sound field described by $\mathbf{B}_{40}$, the matrix of order-40 HOA signals:

$$\mathbf{B}_{40} = \mathbf{B}_{40}^{(\mathrm{dir})} + \mathbf{B}_{40}^{(\mathrm{dif})}\,, \tag{19}$$

where $\mathbf{B}_{40}^{(\mathrm{dir})}$ and $\mathbf{B}_{40}^{(\mathrm{dif})}$ denote the matrices of the HOA signals corresponding to the 6 plane waves and the diffuse background, respectively. The HOA signals are 1024-sample long. As well, the energies of the diffuse and directional components of the sound field are equal, which corresponds to a signal-to-noise ratio of 0 dB if we assume the diffuse background is noise. The ground truth, order-40, energy map for this sound field, calculated using the method described in Section 2, is shown in Figure 1(a).

We now assume that this sound field is measured using an SMA providing the exact HOA signals up to order 2. The order-2 energy map of the sound field is shown in Figure 1(b). This map has limited resolution as expected. Using the method described in Section 3, the order-2 HOA signals were upscaled to order 40 to provide a high resolution map. This map is shown in Figure 1(c). Clearly, the upscaling failed due to the presence of a non-sparse sound field. The IRLS solver found sound sources in directions where there where none. This occurs because spurious plane-waves are used to explain the signals originating from the diffuse background. Lastly, Figure 1(d) shows the composite map obtained using the subspace pre-processing method described in Section 4. Compared to the map presented in Figure 1(c), this map matches the reference map shown in Figure 1(a) much more precisely. Although sources 2 and 4 (see Table 1) were merged into one source, the other plane-wave sources were localized with a reasonable accuracy. As well, the energy of the identified sources is approximately that of the actual sources. Regarding the diffuse part of the sound field, the effect of the subspace pre-processing is clearly visible in that some of the diffuse background energy is missing around the dominant sources. However, the energy of the diffuse background is approximately that of the actual sound field in the directions where there is no dominant plane-wave source.

As the method used for separating the diffuse and directional components of the sound field is closely related to MUSIC or, more specifically, EB-MUSIC [3], we compare the above results with the MUSIC spatial spectrum. The MUSIC spatial spectrum corresponding to the order-2 HOA signals is given by:

$$\phi(u) = \frac{1}{\mathbf{y}_2(u)^{\mathsf{T}}\,\mathbf{V}\,\mathbf{\Psi}\,\mathbf{V}^{\mathsf{T}}\,\mathbf{y}_2(u)}\,, \tag{20}$$

where $\mathbf{V}$ is as defined in Equation (15), and $\mathbf{\Psi}$ is the diagonal matrix given by:

$$\mathbf{\Psi} = \mathrm{diag}\left(\left[\psi_1, \psi_2, \psi_3, ..., \psi_{(L+1)^2}\right]\right)\,, \text{with}\;\; \psi_i = \begin{cases} 0 & \text{for } i \le 6 \\ 1 & \text{otherwise} \end{cases}\,. \tag{21}$$

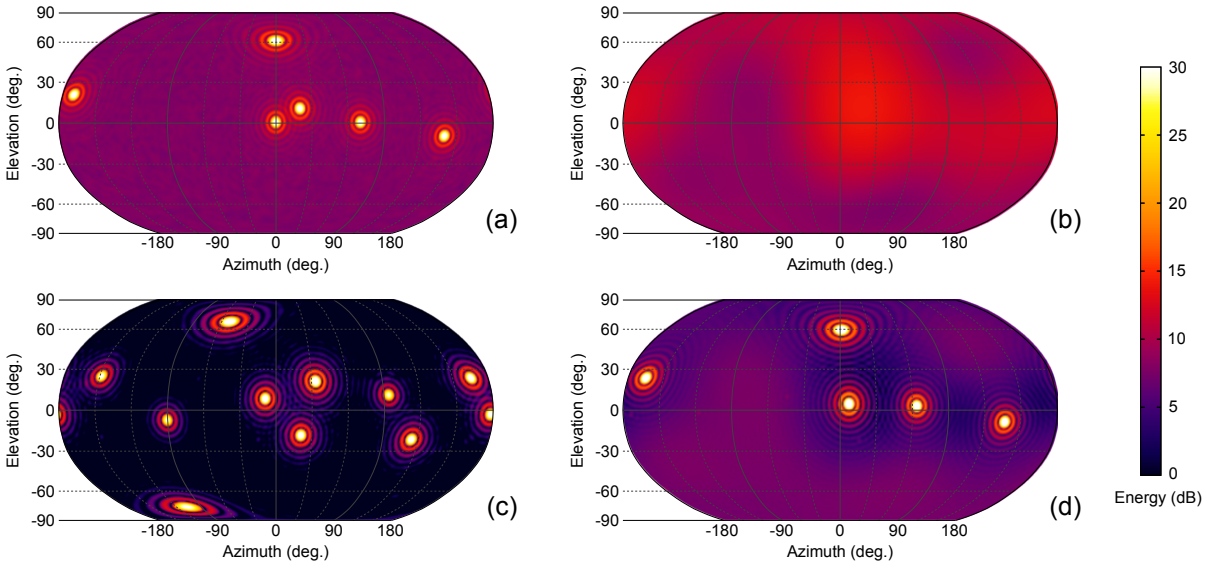

Figure 1: This figure shows the different energy maps obtained in the numerical simulations: a) the true order-40 map; b) the order-2 map; c) the order-40 map obtained from the upscaled HOA signals without pre-processing; d) the composite map obtained using the subspace pre-processing.

Source #	1	2	3	4	5	6
azimuth (°)	-170	0	0	20	70	140
elevation (°)	20	0	60	10	0	-10

Table 1: This table shows the position of the sound sources in the numerical simulation.

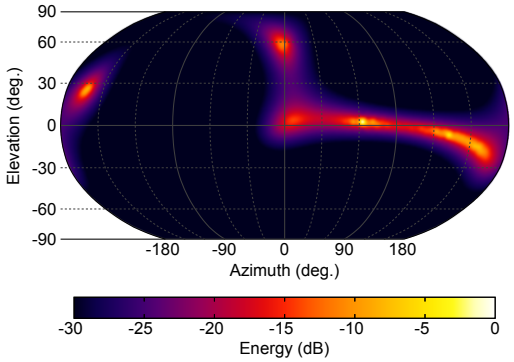

Figure 2: This figure shows the MUSIC spatial spectrum calculated from the order-2 HOA signals.

The MUSIC spatial spectrum is shown in Figure 2. The spectrum underlines a spatial support for the dominant sources. However, with the exceptions of sources 1 and 3, it is not very clear where the sources are. As well, the MUSIC spectrum does not provide quantitative information about the energies of the sources. Lastly, note that this spectrum was calculated assuming we knew 6 dominant sources were present, whereas no assumption on the number of sources was made to obtain the map presented in Figure 1(d).

6 Conclusions

In this tutorial, we briefly reviewed sound field imaging using SMAs. Using traditional beamforming methods, sound field images obtained from SMAs have limited resolution. Higher resolution images can be obtained using SR methods, however such methods are likely to fail in the presence of noise or reverberation. In order to increase the robustness of SR-based sound field analysis, the HOA signals are separated into directional and diffuse components, using a simple subspace method. The simulations show that this pre-processing can dramatically improve the accuracy of the SR sound field analysis in the presence of diffuse noise. As well, the proposed pre-processing technique does not require the number of sources to be known.

7 Appendix

Central to this work is the fact one can record HOA signals using a Spherical Microphone Array (SMA). Obtaining HOA signals from the microphone signals is referred to as *HOA-encoding*. In this section, we derive the formulas used for calculating the order-L HOA encoding filters for an SMA consisting of omnidirectional microphones located on the surface of a rigid sphere.

In the absence of measurement noise, the frequency-domain signal recorded by an omnidirectional microphone located on the surface of a rigid sphere with radius R is given by:

$$x(\vartheta, \varphi, f) = \sum_{l=0}^{\Lambda} \left[i^l \left(j_l(kR) - \frac{j_l'(kR)}{h_l^{(2)'}(kR)} h_l^{(2)}(kR) \right) \sum_{m=-l}^{l} Y_l^m(\vartheta, \varphi)\, b_{lm}(f) \right] , \tag{22}$$

where (ϑ, φ) denotes the azimuth and elevation of the microphone, j_l' denotes the derivative of the spherical Bessel function of degree l and $h_l^{(2)}$ and $h_l^{(2)'}$ denote the degree-l spherical Hankel function of the second kind and its derivative, respectively. In Eq. (22), Λ is chosen sufficiently large to ensure convergence to a given precision.

In the following, we wish to express the mathematics using matrix-vector notation. Therefore, the HOA order, L, of the vectors and matrices in a given equation will be denoted by the equal sign $\overset{(L)}{=}$. For a microphone array with P microphones, the vector of the microphone signals can be expressed as the following matrix-vector product:

$$\mathbf{x}(f) \overset{(\Lambda)}{=} \mathbf{Y}_{\text{mic}}^{\mathsf{T}} \mathbf{W}_{\text{mic}}(f)\, \mathbf{b}(f) , \tag{23}$$

where:

- $(\cdot)^{\mathsf{T}}$ denotes the transpose
- $\mathbf{x}(f)$ is the vector of the frequency-domain microphone signals,

$$\mathbf{x}(f) = [x_1(f),\, x_2(f),\, \ldots,\, x_U(f)]^{\mathsf{T}} , \tag{24}$$

- $\mathbf{Y}_{\text{mic}}$ is the matrix whose coefficients are the spherical harmonic function values up to order Λ corresponding to the microphone positions,

$$\begin{aligned} &\mathbf{Y}_{\text{mic}} \overset{(\Lambda)}{=} [\mathbf{y}_1,\, \mathbf{y}_2,\, \ldots,\, \mathbf{y}_P] , \\ &\text{with } \mathbf{y}_p \overset{(\Lambda)}{=} \big[Y_0^0(\vartheta_p, \varphi_p),\, Y_1^{-1}(\vartheta_p, \varphi_p),\, Y_1^0(\vartheta_p, \varphi_p), \\ &\qquad Y_1^1(\vartheta_p, \varphi_p),\, \ldots,\, Y_\Lambda^\Lambda(\vartheta_p, \varphi_p)\big]^{\mathsf{T}} , \end{aligned} \tag{25}$$

  where (ϑ_p, φ_p) denotes the azimuth and elevation of the p-th microphone,

- $\mathbf{W}_{\text{mic}}(f)$ is the matrix of the modal coefficients up to order Λ,

$$\begin{aligned} &\mathbf{W}_{\text{mic}}(f) \overset{(\Lambda)}{=} \text{diag}(\mathbf{w}_{\text{mic}}(f)) , \\ &\text{with } \mathbf{w}_{\text{mic}}(f) \overset{(\Lambda)}{=} [w_0(kR),\, w_1(kR),\, w_1(kR), \\ &\qquad w_1(kR),\, \ldots,\, w_\Lambda(kR)]^{\mathsf{T}} , \\ &\text{and } w_l(kR) = i^l \left(j_l(kR) - \frac{j_l'(kR)}{h_l^{(2)'}(kR)} h_l^{(2)}(kR) \right) , \end{aligned} \tag{26}$$

  where $\text{diag}(\mathbf{a})$ denotes the diagonal matrix whose elements are the elements of vector $\mathbf{a}$.

- $\mathbf{b}(f)$ is the vector of the frequency-domain HOA signals up to the order Λ,

$$\mathbf{b}(f) \overset{(\Lambda)}{=} [b_{00}(f),\, b_{1-1}(f),\, \ldots,\, b_{\Lambda\Lambda}(f)]^{\mathsf{T}} , \tag{27}$$

The HOA-encoding process consists in solving Eq.(23) for the HOA signals. The least-square solution to Eq.(23) is given by:

$$\hat{\mathbf{b}}(f) \overset{(L)}{=} \mathbf{E}(f)\, \mathbf{x}(f) , \tag{28}$$

where $\mathbf{E}(f)$ is defined as the least-square encoding matrix for the HOA signals up to order L and is given by:

$$\mathbf{E}(f) \overset{(L)}{=} \mathbf{\Omega}_{\text{mic}}(f)\, \text{pinv}\left(\mathbf{Y}_{\text{mic}}^{\mathsf{T}}\right) , \tag{29}$$

where $\text{pinv}(.)$ denotes the Moore-Penrose pseudo inverse, $\mathbf{Y}_{\text{mic}}$ is obtained by truncating $\mathbf{Y}_{\text{mic}}$ at order L, and $\mathbf{\Omega}_{\text{mic}}(f)$ is the Tikhonov-regularised inverse of $\mathbf{W}_{\text{mic}}(f)$,

$$\begin{aligned} &\mathbf{\Omega}_{\text{mic}}(f) \overset{(L)}{=} \text{diag}(\boldsymbol{\omega}_{\text{mic}}(f)) , \\ &\text{with } \boldsymbol{\omega}_{\text{mic}}(f)) = [\omega_0(kR),\, \omega_1(kR),\, \omega_1(kR), \omega_1(kR),\, \ldots,\, \omega_L(kR)]^{\mathsf{T}} , \\ &\text{and } \omega_l(kR) = \frac{w_l^*(kR)}{\|w_l(kR)\|^2 + \beta^2} , \end{aligned} \tag{30}$$

where β is the regularisation coefficient.

The Tikhonov regularisation is used to make the encoding process more robust to measurement noise: at low frequencies the value of $w_l(kR)$ for $l > 1$ are very small, which results in huge encoding matrix coefficients if no regularisation is used. In practice, the time-domain microphone signals are encoded as time-domain order-L HOA signals using a matrix of Finite Impulse Response (FIR) encoding filters, $\mathbf{E}(t)$. The matrix of encoding filters is obtained using the frequency-domain encoding matrices calculated using Eq. (29).

References

[1] B.N. Gover, J.G. Ryan, and M.R. Stinson, "Microphone array measurement system for analysis of directional and spatial variation of sound fields," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 1980–1991, 2002.

[2] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *J. Acoust. Soc. Am.*, vol. 5, no. 118, pp. 3094–3103, 2005.

[3] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, 2010.

[4] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberent environments using EB-ESPRIT with spherical microphone arrays," in *Proceedings of the 2011 ICASSP*, Prague, Czech Republic, May 2011.

[5] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. AP-34, no. 3, pp. 276–280, 1986.

[6] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[7] A. Wabnitz, N. Epain, A. McEwan, and C.T. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *Proceedings of WASPAA*, New Paltz, NY, USA, October 2011.

[8] A. Wabnitz, N. Epain, and C.T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *Proceedings of the 2012 ICASSP*, Kyoto, Japan, March 2012.

[9] P.K.T. Wu, N. Epain, and C.T. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *Proceedings of the 2012 ICASSP*, Kyoto, Japan, March 2012.

[10] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, Academic Press, London, 1999.

[11] N. A. Gumerov and R. Duraiswami, *Fast multipole methods for the Helmholtz equation in three dimensions*, Elsevier, The Netherlands, 2005.

[12] J. Daniel, *Représentation de Champs Acoustiques, Application à la Transmission et à la Reproduction de Scènes Sonores Complexes dans un Contexte Multimédia*, Ph.D. thesis, Université Paris 6, 2000.

[13] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[14] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

Spatial Sound Synthesis with Loudspeakers

Sascha Spors and Franz Zotter

Institute of Communications Engineering
Universität Rostock
Sascha.Spors@uni-rostock.de

Institute of Electronic Music and Acoustics
University of Music and Performing Arts Graz
zotter@iem.at

1 Introduction

The idea of accurately synthesizing a captured sound field by multiple loudspeakers surrounding an extended listening area has been pursued for several decades [1, 2, 3]. It is generally assumed that accurate synthesis of the captured sound scene results in a perceived auditory scene that is imperceptible from the original auditory scene. Techniques which are based on this concept are termed as Sound Field Synthesis (SFS) approaches. The most common representatives are Wave Field Synthesis (WFS) [4] and (near-field compensated) higher-order Ambisonics (HOA) [5].

This lecture gives an overview on the physical basics of SFS that provide a theoretical way of unique and accurate synthesis. However, spatial sampling and acoustical design limit the achievable accuracy. The basics of spatial sampling are briefly reviewed. The systems we can build nowadays are physically sill inaccurate above the mid frequencies, however they do deliver a good perceived quality. Eventually, their good properties are evident in experimental studies.

2 Sound Field Synthesis

2.1 Theoretical Basis

The solution of the homogeneous wave equation for the interior Problem within a bounded source free region V (see Figure 1) with smooth boundary ∂V is provided by the Kirchhoff-Helmholtz integral [6, 7, 8]

$$\oint_{\partial V} \left(P(\mathbf{x}_0, \omega) \, \frac{\partial G(\mathbf{x}|\mathbf{x}_0, \omega)}{\partial(\mathbf{n}, \mathbf{x}_0)} - G(\mathbf{x}|\mathbf{x}_0, \omega) \, \left. \frac{\partial P(\mathbf{x}, \omega)}{\partial(\mathbf{n}, \mathbf{x})} \right|_{\mathbf{x}=\mathbf{x}_0} \right) dS_0 = \begin{cases} P(\mathbf{x}, \omega) & \text{for } \mathbf{x} \in V \\ \frac{1}{2} P(\mathbf{x}, \omega) & \text{for } \mathbf{x} \in \partial V \\ 0 & \text{for } \mathbf{x} \in \bar{V} \end{cases} \tag{1}$$

where $G(\mathbf{x}|\mathbf{x}_0, \omega)$ denotes the Green's function, $P(\mathbf{x}_0, \omega)$ the sound pressure on the boundary ∂V ($\mathbf{x}_0 \in \partial V$), and $\mathbf{n}$ the inward pointing normal vector of ∂V. Equation (1) provides also the solution of the exterior problem when the normal vector $\mathbf{n}$ points outwards and V and $\bar{V}$ are exchanged. The abbreviation $\frac{\partial}{\partial(\mathbf{n},\mathbf{x})}$ denotes the directional gradient, for instance

$$\left. \frac{\partial P(\mathbf{x}, \omega)}{\partial(\mathbf{n}, \mathbf{x})} \right|_{\mathbf{x}=\mathbf{x}_0} = \left. \langle \nabla_{\mathbf{x}} P(\mathbf{x}, \omega), \mathbf{n}(\mathbf{x}) \rangle \right|_{\mathbf{x}=\mathbf{x}_0} , \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product of two vectors. The directional gradient is defined as the gradient of $P(\mathbf{x}, \omega)$ taken with respect to $\mathbf{x}$, projected on the normal vector $\mathbf{n}$ and evaluated at $\mathbf{x}_0$.

The Green's function $G(\mathbf{x}|\mathbf{x}_0, \omega)$ represents the solution of the inhomogeneous wave equation for excitation with a spatial Dirac pulse at $\mathbf{x}_0$. The free-field Green's function $G_0(\mathbf{x}|\mathbf{x}_0, \omega)$ can be interpreted as the spatio-temporal transfer function of a monopole placed at $\mathbf{x}_0$ and its directional gradient as the resective function of a dipole at $\mathbf{x}_0$.

Equation (1) states that the sound field $P(\mathbf{x}, \omega)$ inside V is fully determined by the pressure $P(\mathbf{x}, \omega)$ and its directional gradient on the boundary ∂V. If there is a continuous distribution of monopole and dipole sources (single and double layer potential) on the boundary ∂V, weighted by the gradient of the sound pressure and the sound pressure itself, the sound field within V is fully determined. In the context of SFS, the monopole and dipole sources on the boundary are referred to as (monopole/dipole) *secondary sources*.

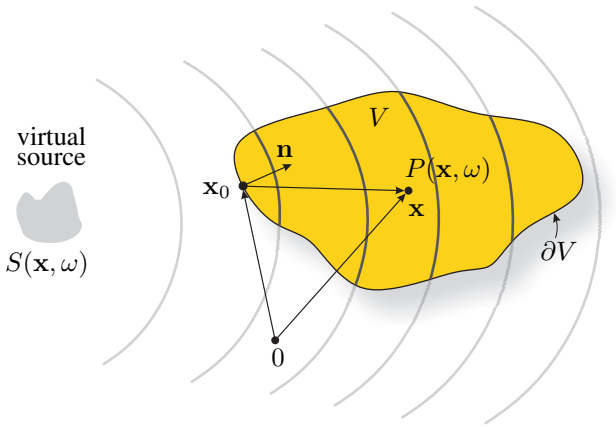

Figure 1: Illustration of the geometry used to discuss the physical fundamentals of sound field synthesis.

For a practical implementation it is desirable to discard one of the two types of secondary sources. Typically a monopole only solution is favorable, since these can be realized reasonably well by loudspeakers with closed cabinets. Hence we wish to find a solution of the shape

$$P(\mathbf{x},\omega) = \oint_{\partial V} D(\mathbf{x}_0,\omega) G(\mathbf{x}|\mathbf{x}_0,\omega) dS_0 \tag{3}$$

where weight (driving function) $D(\mathbf{x}_0,\omega)$ of the secondary sources should be chosen such that $P(\mathbf{x},\omega) = S(\mathbf{x},\omega)$ within V. The question arises if a unique solution can be found?

In order to show that a unique solution indeed exists we follow the simple source or equivalent scattering approach [6, 9]. Lets consider a problem that is equivalent exterior to the one depicted in Figure 1. Contrary to the interior problem in eq. (1), the superposition of the virtual source $S(\mathbf{x},\omega)$ with the sound field inside the Kirchhoff-Helmholtz integral ($\mathbf{x} \in V$) must vanish now

$$S(\mathbf{x},\omega) + \oint_{\partial V} \left(G(\mathbf{x}|\mathbf{x}_0,\omega) \left.\frac{\partial P_t(\mathbf{x},\omega)}{\partial(\mathbf{n},\mathbf{x})}\right|_{\mathbf{x}=\mathbf{x}_0} - P_t(\mathbf{x}_0,\omega) \frac{\partial G(\mathbf{x}|\mathbf{x}_0,\omega)}{\partial(\mathbf{n},\mathbf{x}_0)} \right) dS_0 = 0 \tag{4}$$

Hence, this mental experiment results in a sound free zone within V. When approaching the boundary ∂V from the outside, the total sound field $P_t(\mathbf{x}_0,\omega)$ does not vanish and describes the superposition of the incident sound field $S(\mathbf{x},\omega)$ and the field $P_{\text{sc}}(\mathbf{x},\omega)$ generated by the Kichhoff-Helmholtz integral $P_t(\mathbf{x}_0,\omega) = S(\mathbf{x}_0,\omega) + P_{\text{sc}}(\mathbf{x}_0,\omega)$. Assuming a sound soft boundary ∂V with $P_t(\mathbf{x}_0,\omega) = 0$ simplifies (4) to

$$S(\mathbf{x},\omega) + \oint_{\partial V} \left(G(\mathbf{x}|\mathbf{x}_0,\omega) \left.\frac{\partial P_t(\mathbf{x},\omega)}{\partial(\mathbf{n},\mathbf{x})}\right|_{\mathbf{x}=\mathbf{x}_0} \right) dS_0 = 0 \tag{5}$$

This results holds for the enclosed volume V ($\mathbf{x} \in V$). Hence when the driving function is chosen as

$$D(\mathbf{x_0},\omega) = \left.\frac{\partial(S(\mathbf{x},\omega) + P_{\text{sc}}(\mathbf{x},\omega))}{\partial(\mathbf{n},\mathbf{x})}\right|_{\mathbf{x}=\mathbf{x}_0} \tag{6}$$

the sound field within V cancels exactly with the field of any virtual source $S(\mathbf{x},\omega)$. This equivalent sound-soft scattering problem for the exterior $\bar{V}$ proves that any interior field can be synthesized with opposite sign if the original field $S(\mathbf{x},\omega)$ is removed. The field in the exterior region $\bar{V}$ corresponds to the scattered field $P_{\text{sc}}(\mathbf{x},\omega)$.

For SFS it remains to derive the scattered sound field $P_{\text{sc}}(\mathbf{x},\omega)$ for a given incident field $S(\mathbf{x},\omega)$ and geometry of the boundary ∂V. Unless all relevant constraints imposed by the Kirchhoff-Helmholtz equations (cf. eq. (1)) of the equivalent scattering problem are enforced, the solution may be subject to non-uniqueness issues. However, this has been solved in various different ways. In the literature on the boundary element method (BEM), the CHIEF point method [10] or the Burton-Miller method [11] provide solutions. In particular, some sound fields $S(\mathbf{x},\omega)$ already fulfill the sound-soft boundary condition $S(\mathbf{x}_0,\omega)$ on the surface $\mathbf{x}_0 \in \partial V$ without a scattered field $P_{\text{sc}}(\mathbf{x},\omega)$. It therefore appears as if the scattering problem would degenerate to a trivial problem

$$\oint_{\partial V} D(\mathbf{x}_0,\omega) G(\mathbf{x}|\mathbf{x}_0,\omega) dS_0 = 0 \tag{7}$$

that would already be fulfilled by $D(\mathbf{x}_0, \omega) = 0$. For such a case it is not sufficient to rely on the equation for the total sound pressure on ∂V. To circumvent the trivial case, the CHIEF point method takes into account that the integral of the equivalent scattering problem must vanish inside V. Alternatively, the Burton-Miller method takes into account that the normal gradient of the integral equals half the derivative of the total sound pressure on ∂V.

Overall, it can be concluded from the considerations given so far, that in most situations a unique solution exists for the single layer potential problem given by Equation (3).

2.2 Explicit Solution and Higher-Order Ambisonics

The integral equation (3) constitutes a Fredholm operator of zero index [12, 13, 9]. A solution to (3) is found by expanding its Kernel, which is given by the Green's function, into an orthogonal set of basis functions ψ_n and $\bar{\psi}_n$

$$G(\mathbf{x}|\mathbf{x}_0, \omega) = \sum_{n=1}^{N} \tilde{G}(n, \omega) \left(\bar{\psi}_n(\mathbf{x}_0) \otimes \psi_n(\mathbf{x}) \right) \tag{8}$$

where $\tilde{G}(n, \omega)$ denote the series expansion coefficients. If above expansion together with the expansions of the driving function $D(\mathbf{x}, \omega)$ and virtual source $S(\mathbf{x}, \omega)$ are introduced into (3) one gets

$$\tilde{D}(n, \omega) = \frac{\tilde{S}(n, \omega)}{\tilde{G}(n, \omega)} \tag{9}$$

with

$$D(x, \omega) = \sum_{n=1}^{N} \tilde{D}(n, \omega) \psi_n(\mathbf{x}) \tag{10}$$

The explicit formulation of the basis functions ψ_n and $\bar{\psi}_n$ depends on the geometry of the boundary ∂V and the dimensionality of the problem.

Higher-order Ambisonics assumes spherical secondary source distributions. For spherical boundaries ∂V the basis functions are given by the surface spherical harmonics [6]. Equation (9) can be understood as a generalization of to the mode-matching approach used in the derivation of HOA [5].

The driving function for the synthesis of a plane wave with traveling direction $(\theta_{\text{pw}}, \phi_{\text{pw}})$ using a spherical secondary source distribution with radius R is given as [14]

$$D_{\text{pw}}(\alpha_0, \beta_0, \omega) = \frac{1}{2\pi R} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{4\pi (-i)^n Y_n^m(\theta_{\text{pw}}, \phi_{\text{pw}})^*}{-j \frac{\omega}{c} h_n^{(2)}(\frac{\omega}{c} R)} Y_n^m(\alpha_0, \beta_0) \tag{11}$$

where Y_n^m denote the surface spherical harmonics, $h_n^{(2)}$ the spherical Hankel function of second kind and $x_0 = R \cos\alpha_0 \sin\beta_0$, $y_0 = R \sin\alpha_0 \sin\beta_0$, $z_0 = R \cos\beta_0$.

2.3 High-frequency Approximation and Wave Field Synthesis

In the boundary element method (BEM) a high-frequency formulation of the Kirchhoff-Helmholtz integral (1) is known which is of special interest here. Its derivation and connection to Wave Field Synthesis is briefly outlined in the following.

The explicit form of the Green's function depends on the dimensionalty of the underlying problem. For three-dimensional space it is given as [6]

$$G_{0,3\text{D}}(\mathbf{x}|\mathbf{x}_0, \omega) = \frac{1}{4\pi} \frac{e^{-j \frac{\omega}{c} \|\mathbf{x} - \mathbf{x}_0\|}}{\|\mathbf{x} - \mathbf{x}_0\|} \tag{12}$$

The directional gradient of the Green's function, as required for the Kirchhoff-Helmholtz integral, can be derived as

$$\frac{\partial G_{0,3\text{D}}(\mathbf{x}|\mathbf{x}_0, \omega)}{\partial(\mathbf{n}, \mathbf{x_0})} = \frac{1 + j \frac{\omega}{c} r}{r} \cdot \cos\phi \cdot G_{0,3\text{D}}(\mathbf{x}|\mathbf{x}_0, \omega) \tag{13}$$

where $r = \|\mathbf{x} - \mathbf{x}_0\|$ and $\cos\phi = \frac{\langle \mathbf{x} - \mathbf{x}_0, \mathbf{n} \rangle}{\|\mathbf{x} - \mathbf{x}_0\|}$ with $\phi = \angle(\mathbf{x} - \mathbf{x}_0, \mathbf{n})$. Introducing (13) into (1) and rearranging the terms yields

$$P(\mathbf{x}, \omega) = \oint_{\partial V} \left(P(\mathbf{x}_0, \omega) \cdot \frac{1 + j \frac{\omega}{c} r}{r} \cdot \cos\phi - \left. \frac{\partial P(\mathbf{x}, \omega)}{\partial(\mathbf{n}, \mathbf{x})} \right|_{\mathbf{x}=\mathbf{x}_0} \right) G(\mathbf{x}|\mathbf{x}_0, \omega) \, dS_0 \tag{14}$$

for $\mathbf{x} \in V$. As a result we have achieved a representation of the KHI using weighted secondary monopole sources only. The weight of the secondary sources, given by the terms within the brackets, depends on r and ϕ and hence on the field point $\mathbf{x}$ (listener position). Since we seek for a global solution within V this dependency is not desired. We proceed with the following two assumptions: (i) $(\mathbf{x} - \mathbf{x}_0) || \mathbf{n}$ and (ii) $\frac{\omega}{c} \|\mathbf{x} - \mathbf{x}_0\| \gg 1$. The first assumption results in $\cos\phi = 1$. Introducing this together with the second assumption into (14) yields an approximation where the weight of the Green's function does not depend on the field point $\mathbf{x}$

$$P(\mathbf{x}, \omega) = \oint_{\partial V} \underbrace{\left(j \frac{\omega}{c} P(\mathbf{x}_0, \omega) - \frac{\partial P(\mathbf{x}, \omega)}{\partial (\mathbf{n}, \mathbf{x})} \bigg|_{\mathbf{x}=\mathbf{x}_0} \right)}_{D(\mathbf{x}_0, \omega)} G(\mathbf{x}|\mathbf{x}_0, \omega) \, dS_0 \tag{15}$$

This approximation is known as the high frequency boundary element method [15].

Lets consider the synthesis of a unit amplitude plane wave with traveling direction $\mathbf{n}_{\text{pw}}$

$$S_{\text{pw}}(\mathbf{x}, \omega) = e^{-j \frac{\omega}{c} \langle \mathbf{n}_{\text{pw}}, \mathbf{x} \rangle} \tag{16}$$

Introducing the directional gradient of $S_{\text{pw}}(\mathbf{x}, \omega)$ into the driving function given by (15) yields

$$D_{\text{pw}}(\mathbf{x}_0, \omega) = (1 + \langle \mathbf{n}_{\text{pw}}, \mathbf{n}(\mathbf{x}_0) \rangle) \cdot j \frac{\omega}{c} \cdot e^{-j \frac{\omega}{c} \langle \mathbf{n}_{\text{pw}}, \mathbf{x} \rangle} \tag{17}$$

One can further approximate the terms in the brackets by

$$(1 + \langle \mathbf{n}_{\text{pw}}, \mathbf{n}(\mathbf{x}_0) \rangle) = \begin{cases} 2 \langle \mathbf{n}_{\text{pw}}, \mathbf{n}(\mathbf{x}_0) \rangle & \text{for } \langle \mathbf{n}_{\text{pw}}, \mathbf{n}(\mathbf{x}_0) \rangle > 0 \\ 0 & \text{for } \langle \mathbf{n}_{\text{pw}}, \mathbf{n}(\mathbf{x}_0) \rangle < 0 \end{cases} \tag{18}$$

As a result only the secondary sources for which the first condition holds are active. In BEM this is known as determining the visible elements [15] in WFS as secondary source selection criterion [16].

Introducing this approximation into the driving function (17) yields the driving function of three dimensional Wave Field Synthesis [17]

$$D_{\text{pw,3D}}(\mathbf{x}, \omega) = 2 a_{\text{pw}}(\mathbf{x}_0) \, \langle \mathbf{n}_{\text{pw}}, \mathbf{n}(\mathbf{x}_0) \rangle \, j \frac{\omega}{c} e^{-j \frac{\omega}{c} \langle \mathbf{n}_{\text{pw}}, \mathbf{x} \rangle} \tag{19}$$

Hence Wave Field Synthesis can be regarded as a high frequency approximation of the BEM. In the context of WFS, the assumptions used to derive (15) are similar to the ones used in the derivation of WFS by the stationary phase approximation [18].

2.4 2.5-dimensional Reproduction

Loudspeaker arrays are often arranged within a two-dimensional space, for example as a linear or circular array. From a theoretical point of view, the characteristics of the secondary sources in such setups should conform to the two-dimensional free-field Green's function. Its sound field can be interpreted as the field produced by a line source [6]. Loudspeakers exhibiting the properties of acoustic line sources are not practical. Using point sources as secondary sources for the reproduction in a plane results in a dimensionality mismatch. Therefore such methods are often termed as *2.5-dimensional synthesis* techniques. It is well known from WFS and HOA, that 2.5-dimensional reproduction techniques suffer from artifacts [18, 14]. Amplitude deviations are most prominent. Similar artifacts will also be present in other sound reproduction approaches which aim at correct reproduction in a plane using secondary point sources.

3 Spatial Sampling of Secondary Source Distribution

Practical setups consist of a finite number of spatially discrete secondary sources. This constitutes a spatial sampling process that may lead to spatial sampling artifacts. Current realizations of SFS techniques result in a coarse sampling and sampling artifacts are present for mid to high frequencies. Hence, spatial sampling has to be considered explicitly in the evaluation of SFS techniques.

The sampling process can be modeled as depicted in Figure 2. The driving function $D(\mathbf{x}_0, \omega)$ is sampled at the spatially discrete loudspeaker positions resulting in the sampled driving function $D_{\text{S}}(\mathbf{x}_0, \omega)$. The Green's function $G_0(\mathbf{x} - \mathbf{x}_0, \omega)$ representing the idealized sound field of the loudspeakers is then weighted with the sampled driving function $D_{\text{S}}(\mathbf{x}_0, \omega)$. In the context of classical sampling and reconstruction theory, the Green's function can be interpreted as

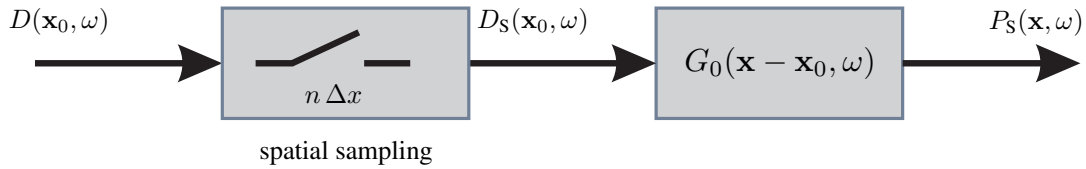

Figure 2: Model of spatial sampling on sound field synthesis.

interpolation filter. In order to investigate the influence of spatial sampling it is convenient to transform the signals into a suitable spectral representation. For circular secondary source distributions this is a Fourier series expansion, for spherical distributions the spherical harmonics expansion. The detailed discussion of spatial sampling artifacts is beyond the scope of this introduction. Results can be found, e. g., in [19, 14, 20, 21, 22]. It is worthwhile noting that the classical sampling theorem, which predicts that no aliasing occurs when at least two sampling points per wavelength are present, does not hold in general for SFS.

Acknowledgments

We would like to thank Frank Schultz for contributions to the derivation of the high-frequency approximation of the Kirchhoff-Helmholtz integral and proofreading.

References

[1] W.B. Snow. Basic principles of stereophonic sound. *IRE Transactions on Audio*, 3:42–53, March 1955.

[2] J. C. Steinberg and W. B. Snow. Symposium on wire transmission of symphonic music and its reproduction in auditory perspective: Physical factors. *Bell Systems Technical Journal*, XIII(2), April 1934.

[3] Maurice Jessel. *Acoustique théorique: propagation et holophonie*. Masson et Cie, Paris, 1973.

[4] A.J. Berkhout. A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36:977–995, December 1988.

[5] J. Daniel. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In *AES 23rd International Conference*, Copenhagen, Denmark, May 2003. Audio Engineering Society (AES).

[6] E.G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[7] D.L. Colton and R. Kress. *Integral Equation Methods in Scattering Theory*. Wiley, New York, 1983.

[8] N.A. Gumerov and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in three Dimensions*. Elsevier, 2004.

[9] F.M. Fazi. *Sound Field Reproduction*. PhD thesis, University of Southampton, 2010.

[10] Harry A. Schenk. Improved integral formulations for acoustic radiation problems. *Journal of the Acoustical Society of America*, 44(1), 1968.

[11] A. J. Burton and G. F. Miller. The application of integral equation methods to the numerical solution of some exterior boundary-value problems. *Proc. R. Soc. London*, 1971.

[12] L.G. Copley. Fundamental results concerning integral representations in acoustic radiation. *Journal of the Acoustical Society of America*, 44(1):28–32, 1968.

[13] Sascha Spors and Jens Ahrens. Towards a theory for arbitrarily shaped sound field reproduction systems. *Journal of the Acoustical Society of America*, 123(5):3930, May 2008.

[14] Jens Ahrens and Sascha Spors. An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions. *Acta Acustica united with Acustica*, 94(6):988–999, December 2008.

[15] D.W. Herrin, F. Martinus, T.W. Wu, and A.F. Seybert. A new look at the high frequency boundary element and rayleigh integral approximations. In *SAE 2003 Noise & Vibration Conference and Exhibition*, Grand Traverse,US, 2003.

[16] Sascha Spors. Extension of an analytic secondary source selection criterion for wave field synthesis. In *123th Convention of the Audio Engineering Society*, New York, USA, October 2007.

[17] Sascha Spors, Rudolf Rabenstein, and Jens Ahrens. The theory of wave field synthesis revisited. In *124th Convention of the Audio Engineering Society*, May 2008.

[18] J.-J. Sonke, D. de Vries, and J. Labeeuw. Variable acoustics by wave field synthesis: A closer look at amplitude effects. In *104th AES Convention*, Amsterdam, Netherlands, May 1998. Audio Engineering Society (AES).

[19] Sascha Spors and Rudolf Rabenstein. Spatial aliasing artifacts produced by linear and circular loudspeaker arrays used for wave field synthesis. In *120th Convention of the Audio Engineering Society*, Paris, France, May 2006.

[20] Sascha Spors and Jens Ahrens. A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling. In *125th Convention of the Audio Engineering Society*, October 2008.

[21] Jens Ahrens and Sascha Spors. A modal analysis of spatial discretization in spherical loudspeaker arrays used for sound field synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 20(9):2564–2574, November 2012.

[22] Sascha Spors and Jens Ahrens. Spatial sampling artifacts of wave field synthesis for the reproduction of virtual point sources. In *126th Convention of the Audio Engineering Society*, May 2009.

Psychoacoustic Experiments with Loudspeaker-Based Virtual Acoustics

Florian Völk
Arbeitsgruppe Technische Akustik
Prof. Dr.-Ing. Hugo Fastl
Institute for Human-Machine-Interaction
Technische Universität München

1 Introduction

This lecture discusses some methodical aspects and results of psychoacoustic experiments with loudspeaker-based virtual acoustics (VA). Experiments of this kind are typically employed for perceptually evaluating the performance of spatial audio systems. VA systems can also help rising and solving questions regarding properties of the hearing system (e. g. Völk and Fastl 2011a). While the material presented on these pages is sufficient for the course, some references are given, directing the interested reader to additional information.

2 Methods

The fundamental concern of Psychoacoustics is establishing relationships between physical stimulus properties and corresponding hearing sensations. Traditionally, a frontally incident plane wave propagating in a free sound field, described by closed mathematical equations, is regarded as the most simple stimulus for psychoacoustic studies (Fastl and Zwicker 2007). For basic studies on directional hearing, more complex scenarios consisting of one or more point sources are typically used (Blauert 1997). Consequently, the stimulus (the spatio-temporal sound pressure field) becomes more complicated. In every case, the physical properties of the considered stimuli (in directional hearing especially the source positions and extensions) are mathematically clearly describable, which allows for the definition of psychoacoustic relationships.

2.1 Psychoacoustic Experiments in Virtual Acoustics

When dealing with VA, the situation becomes more complicated. For VA rendering, physically existing secondary sound sources are used to create the sound field that would emerge from one or more virtual (physically not existing) primary sources (cf. Völk 2011, Völk and Fastl 2012). Conducting a psychoacoustic experiment in VA is usually done by varying a (virtual) physical property of the physically not present primary source by adjusting the rendering algorithm, and by aiming at inspecting related changes of hearing sensations. In other words, the virtual primary source's properties are regarded as the stimulus magnitudes (e. g. Völk et al. 2010).

This procedure is insofar in line with conventional psychoacoustic experiments, as relations between physical stimulus parameters and corresponding hearing sensations are addressed. However, a major difference to conventional experiments is that the stimuli are created indirectly by the VA system. This fact results in a considerable restriction of generality, since the specific VA system's characteristics (especially signal processing and hardware influences) are inevitably included in the results. For that reason, in addition to the conventional description of the experimental

setup, detailed knowledge about the synthesis algorithm (including its implementation) and about the complete electro-acoustic signal processing chain is necessary to allow for the meaningful discussion of the results of psychoacoustic experiments in VA.

One might argue that psychoacoustic studies on VA are redundant, since the aim of VA procedures is synthesizing the sound field created by one or more primary sources, and the psychoacoustic data known for real sources can be applied to a virtual scenario. This argumentation is not valid if the VA system considered does not reproduce the pressure field of the reference scene completely correct. Especially in reflective environments, listening room influences typically disturb the synthesized wave field. In these cases, it is not necessarily clear that the auditory perceptions evoked by the secondary sources equal those evoked by the primary field, and psychoacoustic relationships determined for real (physically present) sound sources must not be applied to VA scenarios without verifying their validity for the specific setup (Völk 2010).

2.2 Minimum Audible Angle Measurement

The angle between the lines from the center of the head to two stationary sound sources at the same distance with just noticeably different positions when sounded in succession is referred to as the minimum audible angle (MAA). For real sound source, Mills (1958) measured MAAs in an anechoic environment. His results indicate the MAA to depend on the sound incidence direction and the spectral stimulus content, where minimal MAAs below 1° occur for frontal sound incidence. Following Perrott and Pacheco (1989), an adaptive two-alternative forced choice (2-AFC) 2-down/1-up method is typically used for MAA measurements. Therefore, the stimuli to be compared are presented by (real or virtual) sound sources at the same distance under head-related angles symmetric around the reference direction, and the step size is adapted by Parameter Estimation by Sequential Testing (PEST). It is the subjects' task to indicate by pressing one of two buttons where the second hearing sensation occurred with regard to the first hearing sensation (in a specified plane). The presentation sequence is chosen randomly and the procedure is repeated until both, the deviation between the last two minimum and the deviation between the last two maximum values are below a threshold value. Since the 2-down/1-up method converges to the 70.7% point of the psychometric function, the MAA is defined as the angular threshold where about 71% of all relative position judgments are correct. The adaptive procedure is repeated three times per stimulus and subject, and the intra-individual median per stimulus is taken as the individual result (cf. Völk and Fastl 2011b).

2.3 Minimum Audible Distance Measurement

The distance between two stationary sound sources on a line through the center of the head with just noticeably different positions when sounded in succession is referred to as the minimum audible distance (MAD). For real sources, Zahorik et al. (2005) reports the accuracy of distance judgments to decrease with the source distance. Similar to the MAA measurement, a two-alternative forced choice 2-down/1-up method combined with PEST for the step size adaption is typically used for assessing the MAD. The stimuli to be compared are positioned at different distances symmetrical around a reference distance. The subjects are asked to indicate where the second hearing sensation occurred in relation to the first (in a specified plane) by pressing one of two buttons, while the presentation sequence is chosen randomly.

2.4 Stimuli

Basic experiments targeting only the auditory modality should be conducted with the subject seated in a darkened laboratory, aiming at applying the visual stimulus darkness and therefore providing a controlled situation with regard to audio-visual interactions. As sound stimuli, broadband (20 Hz to 20 kHz) uniform exciting noise (UEN) impulses (Fastl and Zwicker 2007) are used often, for containing equal intensity in all critical bands and thus assumed to provide the listener with all spectral localization cues at the same perceptual weight. To provide dynamic localization cues, 700 ms impulse duration, 20 ms Gaussian gating, and 300 ms pause between the two impulses are suited. Low-pass, high-pass, or band-pass filtered versions are used if frequency dependencies are to be studied.

3 Results

The MAA and MAD measurements shown here were conducted in a darkened, reverberant laboratory (6.8 m × 3.9 m × 3.3 m) at the Institute for Human-Machine-Interaction of Technische Universität München. The lab's average reverberation time can be varied between 50 ms (low) and 140 ms (high). The results are shown as the medians and inter-quartile ranges of the individual means over three repetitions per condition. The (virtual) sound sources were plane and spherical primary wave fronts generated by wave field synthesis (WFS). Details on the setup and signal processing are given by Völk (2010) and Völk and Fastl (2012), respectively.

3.1 Minimum Audible Angle Results

The MAA experiments were carried out by four experienced subjects. Discussed are influences of the primary source characteristics and position, the sound incidence direction, the listening environment/reverberation time, and the spectral stimulus content (cf. Völk and Fastl 2011b).

3.1.1 Primary Source Characteristics and Distance

Figure 1 shows MAAs for broadband UEN impulse pairs, radiated from primary point sources at different distances (circles). Left-pointing triangles show the MAAs for low-pass UEN (20 Hz to 2 kHz). Filled symbols indicate results for the lower reverberation time, open symbols for the higher reverberation time. The dashed lines indicate the maxima of the MAA quartiles measured

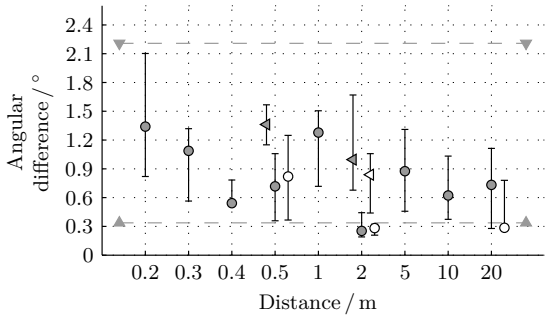

Figure 1: Inter-individual quartiles of minimum audible angles, individually averaged over three repetitions per condition, measured in a specific wave field synthesis setup with uniform exciting noise impulse pairs in a reverberant environment. Frontally positioned point sources, broadband (○) and low-pass noise (◁) for two different room conditions (filled symbols 50 ms, open symbols 140 ms average reverberation time). Dashed lines indicate the maximum and minimum quartiles for plane waves at different levels.

with plane waves at different levels. The results reflect the data known for real sources in their global magnitude. Other than that, no systematic differences arise, not between spherical and plane primary sources, and not for primary point sources at different distances.

3.1.2 Sound Incidence Direction

Figure 2 shows the MAAs for primary point sources synthesized at 2 m distance in different directions. Circles indicate broadband noise stimuli, leftwards pointing triangles low-pass noise stimuli (20 Hz to 2 kHz), and rightward pointing triangles high-pass noise stimuli (2 kHz to 20 kHz). Additionally, results for broadband noise impulses presented by primary plane waves are indicated by horizontal bars. The tendency of a reduced directional resolution for lateral sound incidence

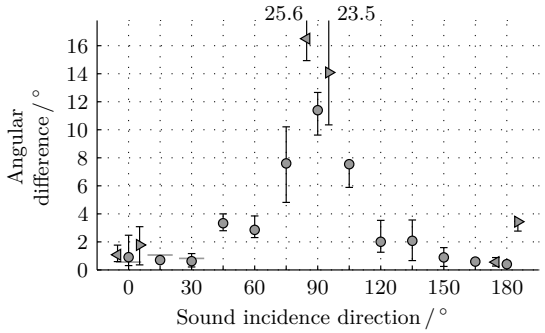

Figure 2: Inter-individual quartiles of minimum audible angles, individually averaged over three repetitions per condition, measured in a specific wave field synthesis setup with uniform exciting noise impulse pairs in a reverberant environment. Depicted are results for primary point sources at 2 m distance for different directions of the sound incidence. Broadband (○), low-pass (◁), and high-pass noise impulse pairs (▷). Horizontal lines indicate the medians for primary plane wave fronts with broadband stimuli.

reported by Mills (1958) is confirmed for WFS sources. Values in the range of 1° are reached for frontal and rearward sound incidence, also comparable to real sources.

3.1.3 Spectral Stimulus Content and Listening Environment

In addition to the results for broadband stimuli discussed so far, high-pass (▷) and low-pass (◁) noise stimuli were included in the experiments shown above (2 kHz lower/higher limiting frequency). Figure 1 indicates for frontally positioned primary point sources at 0.5 m and 2 m distance an increased MAA for stimuli limited to spectral contents below 2 kHz, which is confirmed in figure 2 for frontal and rearward, and especially for lateral sources at 2 m distance. Similar tendencies are visible in figure 2 for stimuli restricted to contents above 2 kHz, where for rearward sound incidence the increase is higher than for low-pass stimuli. The latter fact may indicate that spectral content above 2 kHz contributes more to distinguishing rearward sound incidence directions than spectral content below 2 kHz.

Figure 3 shows data for frontally incident primary plane waves and narrow-band noise stimuli. The light gray contour indicates the results for pure-tone stimuli and real sources of Mills (1958). Globally, the real source data given for anechoic environments by Mills (1958) are confirmed with WFS in a reverberant environment regarding magnitude and frequency dependence.

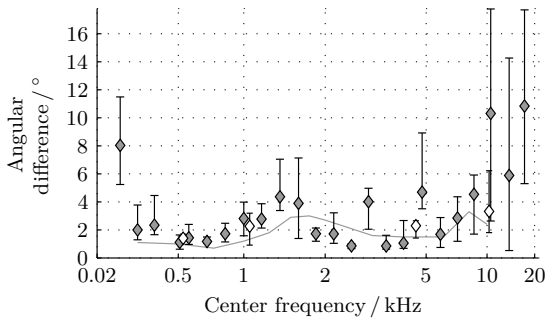

Figure 3: Inter-individual quartiles of minimum audible angles, individually averaged over three repetitions per condition, measured in a specific wave field synthesis setup in a reverberant environment. Depicted are results for frontally incident primary plane waves and critical-band wide narrow-band noise impulse pairs in two different acoustical environments (filled symbols 50 ms, open symbols 140 ms average reverberation time). The gray contour indicates the data for tone impulses and real sources of Mills (1958).

Additionally, figures 1 and 3 contain results for the two different reverberation times 50 ms and 140 ms. In all considered conditions, the reverberation time shows negligible influence on the resulting MAAs.

3.2 Minimum Audible Distance Results

The MAD experiments were carried out by eight experienced subjects (cf. Völk et al. 2012). Figure 4 represents results for frontally positioned primary point sources at different distances and broadband (○) as well as low-pass (20 Hz to 2 kHz, ◁) UEN.

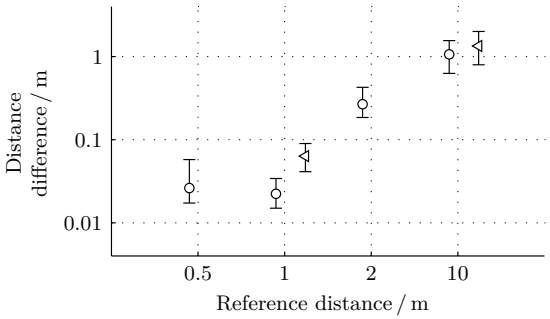

Figure 4: Inter-individual quartiles of minimum audible distances, individually averaged over three repetitions per condition, measured in a specific wave field synthesis setup in a reverberant environment. Depicted are results for frontally positioned primary point sources with broadband (○) and low-pass (◁) uniform exciting noise impulses.

The data confirm the decaying accuracy of distance discrimination with distance reported for real sources by Zahorik et al. (2005). Low-pass filtering results in increased MADs at 1 m reference distance. Separate one factorial analysis of variance indicates for 1 m reference distance a significant main effect of the stimulus [F(1,7)= 6.41; p= 0.0391], but not at 10 m [F(1,7)= 1.79; p= 0.2231].

4 Conclusions

The results presented here show that psychoacoustic experiments for the performance evaluation of virtual acoustics systems are feasible if the virtual source is regarded as the stimulus. For minimum audible angle and minimum audible distance, the relations between (virtual) stimulus parameters and resulting hearing sensations closely resemble the corresponding real situations.

References

Blauert J.: *Spatial Hearing – The Psychophysics of Human Sound Localization.* Revised Edition (The MIT Press, Cambridge, Massachusetts, London, 1997)

Fastl H., E. Zwicker: *Psychoacoustics – Facts and Models.* 3rd Edition (Springer, Berlin, Heidelberg, 2007)

Mills A. W.: On the Minimum Audible Angle. J. Acoust. Soc. Am. **30**, 237–246 (1958)

Perrott D. R., S. Pacheco: Minimum audible angle thresholds for broadband noise as a function of the delay between the onset of the lead and lag signals. J. Acoust. Soc. Am. **85**, 2669–2672 (1989)

Völk F.: Psychoakustische Experimente zur Distanz mittels Wellenfeldsynthese erzeugter Hörereignisse (Psychoacoustic experiments on the distance of auditory events in wave field synthesis). In *Fortschritte der Akustik, DAGA 2010*, 1065–1066 (Dt. Gesell. für Akustik e. V., Berlin, 2010)

Völk F.: System Theory of Binaural Synthesis. In *131st AES Convention* (2011) (Convention Paper 8568)

Völk F., H. Fastl: Locating the Missing 6 dB by Loudness Calibration of Binaural Synthesis. In *131st AES Convention* (2011a) (Convention Paper 8488)

Völk F., H. Fastl: Richtungsunterschiedsschwellen (Minimum Audible Angles) für ein zirkulares Wellenfeldsynthesesystem in reflexionsbehafteter Umgebung (Minimum audible angles for a circular wave field synthesis system in a reverberant environment). In *Fortschritte der Akustik, DAGA 2011*, 945–946 (Dt. Gesell. für Akustik e. V., Berlin, 2011b)

Völk F., H. Fastl: Wave Field Synthesis with Primary Source Correction: Theory, Simulation Results, and Comparison to Earlier Approaches. In *133rd AES Convention* (2012) (Convention Paper 8717)

Völk F., U. Mühlbauer, H. Fastl: Minimum Audible Distance (MAD) by the Example of Wave Field Synthesis. In *Fortschritte der Akustik, DAGA 2012*, 319–320 (Dt. Gesell. für Akustik e. V., Berlin, 2012)

Völk F., M. Straubinger, H. Fastl: Psychoacoustical experiments on loudness perception in wave field synthesis. In *20th Int. Congress on Acoustics (ICA)* (2010)

Zahorik P., D. S. Brungart, A. W. Bronkhorst: Auditory Distance Perception in Humans: A Summary of Past and Present Research. Acta Acustica united with Acustica **91**, 409–420 (2005)