Analysis of the classical singing voice with objective parameters

Voice Directivity and Phonation Mode Analysis

DOCTORAL THESIS

submitted by

Dipl.-Ing. Manuel Brandner

(matriculation number: 00530723)

submitted to

University of Music and Performing Arts Graz
PhD Program (Sound and Music Computing, V 094 750)

Supervisors

Prof. Dr. Alois Sontacchi
Prof. Dr. Robert Höldrich

External Reviewer

Dr. Brian F.G. Katz

Graz, April 13, 2023

Abstract

Although a great deal of research interest has been directed towards the performance analysis of the classical singing voice in recent decades, there is still no commonly used assistance system in higher music education. Acoustic descriptors to describe voice quality, vowel identity and overall efficiency are still investigated or searched for. These descriptors could provide viable feedback for singers during training.

Since the number of features that can be extracted from audio recordings of the classical singing voice is very large, the current thesis focuses on the investigation of the singing voice directivity and, secondly, phonation mode analysis in order to find new objective parameters. Currently available data on both topics are limited. Therefore, the focus is also upon measurement and collection of data for classical singing voice analysis.

A new measurement setup was developed to investigate the properties of singing voice directivity in high detail. The measurement data were compared to simulation data to better understand the high complexity of the voice directivity data. The results show that the mouth opening produces the most noticeable changes, but high variability in the data is also found across different singers. It is also shown that directivity characteristics can be used to distinguish between front and back vowels.

A study was conducted on the perception of changes in the directivity pattern of the voice in a virtual environment. In this way, it is possible to assess whether the differences in directivity found in the measured voice directivity data are relevant to auditory perception. The results showed that minimum changes in a magnitude of 4% are already perceivable for noise, but larger changes of 32% are necessary for speech to be distinguishable to listeners. However, in the measurement data for ten classical singers, a change at this magnitude of 32% is rarely found in the averaged data at certain frequencies. This suggests a limited relevance of changes in voice directivity patterns for different vowels for auditory perception of the singing voice.

The analysis of phonation modes with common signal processing methods is studied and compared to a new approach where modulation characteristics are extracted by using the modulation power spectrum. The results indicate an increased performance for the new developed features to assess phonation modes compared to previously presented features in literature. Furthermore, a design approach for a training tool for phonation mode analysis and vowel identification is presented.

Kurzfassung

Obwohl in den letzten Jahrzehnten ein großes Forschungsinteresse auf die Analyse der klassischen Gesangsstimme gerichtet wurde, gibt es immer noch keine allgemein verwendete computergestütze Stimmanalyse im tertiären Sektor der Gesangsausbildung. Akustische Deskriptoren zur Beschreibung der Stimmqualität, der Vokalidentität und der Gesamteffizienz werden immer noch beforscht und gesucht. Diese Deskriptoren in Verwendung von geeigneter Software könnten Sängerinnen und Sängern während ihrer Ausbildung eine nützliche Unterstützung für das Gesangstraining bieten.

Da die Anzahl der Merkmale, die aus Audioaufnahmen der klassischen Gesangsstimme extrahiert werden können, sehr groß ist, konzentriert sich die vorliegende Arbeit auf die Untersuchung der Richtwirkung der Gesangsstimme und zweitens auf der Stimmqualitätsanalyse, um neue objektive Parameter für den Gesang zu finden. Die derzeit verfügbaren Daten zu beiden Forschungsthemen sind begrenzt. Daher richtet sich auch ein Fokus dieser Arbeit auf die Erfassung und Messung von Daten für die klassische Gesangsstimmanalyse.

Im Zuge dieser Arbeit wurde ein neuer Messaufbau entwickelt, welcher es ermöglicht die Unterschiede in der Richtwirkung von Gesangsstimmen im Detail zu untersuchen. Die Messdaten wurden mit Simulationsdaten verglichen, um die hohe Komplexität der Daten zur Richtwirkung besser zu verstehen. Die Ergebnisse zeigen, dass die Mundöffnung am stärksten Einfluss auf die Richtwirkung hat, aber auch eine hohe Variabilität in der Richtwirkung bei verschiedenen Sängerinnen und Sängern ist festzustellen. Es konnte gezeigt werden, dass die Deskriptoren der Richtcharakteristik zur Unterscheidung zwischen vorderen und hinteren Vokalen verwendet werden können.

Weiters wurden Experimente zur Wahrnehmung von Änderungen in Richtwirkungsmuster der Stimme mit Hilfe von Simulationen und Hörversuchen durchgeführt. Auf diese Weise lässt sich beurteilen, ob die in den Messdaten gefundenen Unterschiede in der Richtwirkung für die auditive Wahrnehmung relevant sind. Die Ergebnisse zeigten, dass bereits minimale relative Änderungen der Richtwirkung in der Größenordnung von 4% bei Rauschsignalen wahrnehmbar sind, aber größere Änderungen von ca. 32% notwendig sind, damit bei Sprache für den Hörer Unterschiede hörbar werden. In den für diese Arbeit erhobenen Messdaten von zehn klassischen Sängerinnen und Sängern finden sich Änderungen in der Größenordnung von 32% nur kaum. Dies deutet auf eine geringere Relevanz der unterschiedlichen Richtwirkung von verschiedenen Vokalen in Bezug auf die auditive Wahrnehmung hin.

Die Analyse von Phonationsmodi (Stimmqualitäten) mit gängigen Signalverarbeitungsmethoden wurde untersucht und mit einem neuen Ansatz verglichen, bei dem Modulationscharakteristika aus dem Modulationsleistungsspektrum extrahiert werden. Die Ergebnisse zeigen, dass die neu entwickelten objektiven Parameter zur Bewertung von Phonationsmodi im Vergleich zu den bisher in der Literatur vorgestellten Merkmalen eine bessere Unterscheidbarkeit aufweisen. Außerdem wird in dieser Arbeit ein Designansatz für ein Trainingstool zur Stimmqualitätsanalyse und Vokalidentifikation im Gesang vorgestellt.

Acknowledgements

I would like to thank all my colleagues for the fruitful discussions, for the all the help and the many coffee breaks during this journey.

I want to express my gratitude to the Institute, Alois Sontacchi, and Robert Höldrich for giving me the chance and encouragement to work on this particular subject. I want to thank Brian Katz for his guidance from afar.

I want to thank my family and especially my wife Klaudia for all the encouraging words, her patience and for all the help.

I would like to dedicate this work to my son Marlon. You shall find what you are looking for in life.

Erklärung

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

Graz, am	(Unterschrift)

Contents

1 Introduction

Although a great deal of research interest has been directed towards the performance analysis of the classical singing voice in recent decades, there is still no commonly used assistance system in higher music education. Acoustic descriptors to describe voice quality, vowel identity and overall efficiency are still investigated or searched for. These descriptors could provide viable feedback for singers during training.

This thesis examines the singing voice directivity and the derivation of characteristics from it as a starting point. Voice directivity explains the sound propagation from the mouth opening and nasal orifices into space. Voice radiation can be characterized by metrics calculated from measured sound pressure levels at defined distances around a defined center position. Research has shown that speech has, to some extent, unique radiation characteristics for different vowels. Furthermore, the size and shape of the mouth opening in classical singing is reported to be a descriptor of vocal efficiency. It has been also shown that directivity affects the perceived distance between the listener and a sound source, thus affecting auditory perception. This work investigates whether simplified acoustic descriptors can be used as objective parameters for the analysis of classical singing. The findings can be used to discuss the potential uses for computer-assisted singing training. It was also investigated whether the differences found in the measured data could result in a change in the perceived sound. Investigating the effects of voice directivity on auditory perception provides information on the effects on the acoustics and self-perception of the voice when singing in concert spaces.

In contrast to the study of the directional characteristics of the singing voice, the study of signal processing methods to distinguish phonation modes in classical singing could help to develop appropriate feedback tools and inform singers about the control and awareness of their own voice. Existing approaches in speech attempt to extract characteristics about the vocal fold movement from the audio signal. Similar information, provided for singers in computer-assisted voice training, could help singing students to maintain a healthy voice throughout the course of their studies. For the present work, common approaches of glottal inverse filtering and cepstral analysis were investigated. In this thesis, a new approach based on cepstral analysis to extract modulation characteristics in classical singing is presented. In order to investigate different phonation modes with current and new signal processing methods in classical singing, the largest currently available dataset for classical singing has been created. The data was analyzed with common and new signal processing methods using a support vector machine. In addition, a listening assessment was conducted to evaluate the newly created data.

Organization Chapter 2 presents the work on the measurement setup, tools and metrics to discuss voice directivity characteristics and investigate the effects on voice directivity in singing due to different mouth openings, vowels, and phonation modes. In the course of the study on voice directivity also simulations of various vocal tract configurations on a rigid sphere have been compared to real measurement data. The data reveals a clear dependence of the mouth opening in singing on the voice directivity characteristics. The found differences are subtle, but can be used to identify front and back vowels. Voice directivity seems to play a subordinate role in auditory perception.

In Chapter 3, the perception of voice in regard to voice directivity in virtual environments is investigated. The virtualization allows us to understand the role of voice directivity more easily. The room acoustics, the spectral composition of sound and the position of the listener in the room play a major role on how sound is perceived.

The second focus of this thesis, the analysis of phonation modes in classical singing, is presented in Chapter 4. New acoustic descriptors for phonation mode analysis are presented, compared to commonly used reference feature sets, and it is shown how they can be used to classify phonation modes by using a support vector machine. Furthermore, the design process of VST-plugin for phonation mode analysis and vowel identification is presented.

Finally, the concluding remarks are presented in Chapter 5.

List of Publications

This thesis consists of an introduction and the publications which are listed here. Please note that the author's contributions to each publication are given at the beginning of the corresponding Sections.

[P1] M. Brandner, M. Frank, and D. Rudrich. DirPat - Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers. *144th AES Convention, Audio Engineering Society Convention e-Brief 425*, Milano, 2018.

[P2] M. Brandner, M. Frank, and A. Sontacchi. A Voice Directivity Measurement System with Facial Tracking and Augmented Acoustics. *Proceedings of the DAGA*, 47:71–74, Vienna, 2021.

[P3] M. Brandner, A. Sontacchi, and M. Frank. Real-Time Calculation of Frequency-Dependent Directivity Indexes in Singing. *Proceedings of the DAGA*, 45:70–73, Rostock, 2018.

[P4] R. Blandin, M. Brandner. Influence of the Vocal Tract on Voice Directivity. *Proceedings of the 23rd International Congress on Acoustics*, 23:1795–1801, Aachen, 2019.

[P5] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi. A Pilot Study on the Influence of Mouth Configuration and Torso on Singing Voice Directivity. *Journal Acoustical Society of America*, 148(3):1169–1180, 2020. doi:10.1121/10.0001736.

[P6] R. Blandin, B. Monson, and M. Brandner. Influence of Speech Sound Spectrum on the Computation of Octave Band Directivity Patterns. *Proceedings of the FA2020 Conference*, 2027–2033, 2020. doi:10.48465/fa.2020.0446.

[P7] M. Brandner, M. Frank, and A. Sontacchi. Horizontal and Vertical Voice Directivity Characteristics of Sung Vowels in Classical Singing. *MDPI Acoustics*, 4:849–866, 2022. doi:10.3390/acoustics4040051.

[P8] M. Frank and M. Brandner. Perceptual Evaluation of Spatial Resolution in Directivity Patterns. *Proceedings of the DAGA*, 46:74–77, Vienna, 2019.

[P9] M. Frank and M. Brandner. Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions. *Proceedings of ICSA*, 5:131–135, Ilmenau, 2019, doi:10.22032/dbt.39966.

[P10] M. Brandner, P. A. Bereuter, S. R. Kadiri, A. Sontacchi. Classification of Phonation Modes in Classical Singing using Modulation Power Spectral Features. *in IEEE Access*, 11:29149–29161, 2023. doi:10.1109/ACCESS.2023.3260187.

[P11] P. A. Bereuter, F. Kraxberger, M. Brandner and A. Sontacchi. Design of a Vowel and Voice Quality Indication Tool Based on Synthesized Vocal Signals. *150th AES Convention, Audio Engineering Society Convention e-Brief 642*, 2021.

2 Voice Directivity Analysis

Several studies in literature investigated the voice directivity patterns of human speech, but only a few investigated singing. No research could clearly reveal to which extent a person can control or at least incidentally change these patterns dependent on the phoneme or specific singing strategy. In order to quantify how much voice directivity changes in classical singing are due to the effects of changes in mouth opening, vowels, or even phonation mode, a suitable measurement setup and methodology must first be established.

In Section 2.1 the measurement setup and initial tools to discuss voice directivity are presented. The initial tools were used to investigate general questions about the effects on voice directivity, appropriate measurement procedures, precision of measurements with real singers. The enhanced measurement setup is presented in Section 2.2, which includes additional sensors to give information about the effective mouth opening used by the singers. In addition, augmented acoustics are used to provide a better recording environment for the singers. The work presented in Section 2.3 investigates whether real-time calculations can meet the quality of off-line measurement data.

Sections 2.4 and 2.5 investigate the effects of the vocal tract configuration and mouth opening on voice directivity. The results show a dependency on the size and shape of the body and the mouth opening. Higher order propagation modes due to specific vocal tract geometries seen in simulations play a subordinate role in the analysis of voice directivity characteristics in classical singing. The role of the type of analysis in terms of frequency resolution, for example, and investigations on larger datasets are presented in Sections 2.6 and 2.7.

2.1 Measurement Setup and Tools to analyze Source Directivity

This work was published as:

M. Brandner, M. Frank, and D. Rudrich. DirPat - Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers. *144th AES Convention, Audio Engineering Society Convention e-Brief 425*, Milano. 2018.

The idea and concept of this article were outlined by me, the first author, with help from the second author. I wrote the original draft of the manuscript with periodical contributions from the third and second author. The revision and editing was done by me with help from the third and second author. I did all of the programming, and prepared the samples for the public available database.

Audio Engineering Society

Convention e-Brief 425

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

DirPat - Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers

Manuel Brandner[1], Matthias Frank[1], and Daniel Rudrich[1]

[1] *University of Music and Performing Arts Graz, Institute of Electronic Music and Acoustics*

Correspondence should be addressed to Manuel Brandner (brandner@iem.at)

ABSTRACT

A measurement repository (DirPat) has been set up to archive all 3D and 2D directivity patterns measured at the Institute of Electronic Music and Acoustics, University of Music and Performing Arts in Graz. Directivity measurements have been made of various loudspeakers, microphones, and also of human speakers/singers for specific phonemes. The repository holds time domain impulse responses for each direction of the radiating or incident sound path. The data can be visualized with the provided 2D and 3D visualization scripts programmed in MATLAB. The repository is used for ongoing scientific research in the field of directivity evaluation of sources or receivers regarding localization, auditory perception, and room acoustic modeling.

1 Introduction

This contribution is in the spirit of open data and reproducible research. It offers source and receiver directivity data measured at the Institute of Electronic Music and Acoustics (IEM) for download from the persistent data repository[1] of the University of Music and Performing Arts, Graz. For visualization and analysis of the data that is stored in AES69 SOFA format (Spatially Oriented Format for Acoustics [1]), directivity inspection tools developed in the last decade at IEM are also included in the repository.

The database of the IEM is comparable to the database presented in [2], however with a denser measurement grid over a full 3D sphere for the loudspeakers and microphones. For human speakers/singers, 2D directivity data for the horizontal and vertical axis is provided. Acquiring detailed spherical measurement data of human speakers/singers and musical instruments in 3D still remains a time consuming process and is not easy to perform without the commitment of the performers. That data will be added to the DirPat repository in the future. For musical instruments a database is presented in Shabtai et al. [3].

This document is outlined as follows: first, the measurement setups for sources and receivers are described, followed by the necessary signal processing of the recorded audio for each scenario. Subsequently, the directivity viewers are introduced, which can be used to visualize the measured directivity data. Finally, exemplary measurement data is presented:

(1) an AKG C414 large-diaphragm condenser microphone with its selectable directivity patterns,
(2) a combination of two guitar cabinets creating a cardioid radiation pattern,
(3) comparison of two human speakers/singers.

[1] https://phaidra.kug.ac.at

2 Measurement of Directivity Patterns

2.1 Measurement Setups

2.1.1 Circular Loudspeaker Array

Impulse response measurements of receivers employ a vertical semicircle of 16 loudspeakers at a radius of 1.5 m with an angular spacing of 11.25°. The used loudspeakers have 3-inch drivers and a wooden rationally symmetric housing [4]. The measurement signal is amplified with a Bittner 8X 100 multichannel amplifier.

2.1.2 Double Circle Microphone Array

For measuring source radiation patterns, a microphone array consisting of two circular rings is employed, cf. Fig. 1. Each of the rings, one placed in the horizontal, the other one in the vertical plane, respectively, can hold up to 32 microphones. The maximum number of microphones is 62 as both rings intersect in the front and back of the array. The angular spacing of the microphones is 11.25° with a radius of 1 m. The apparatus' diameter spans 2.56 m at its widest point. The center-facing side of the rings with a thickness of 21 mm are beveled to reduce reflections. With the setup being held by adjustable loudspeaker stands, the center of the array can be lifted to any height between 1.3 m and 2 m. The microphones mounted to the array are NTI MA 2230 and they are connected to an Andiamo.MC Directout Technologies microphone preamplifier.

2.2 Measurement Procedures

2.2.1 Measurement of Receivers

To measure the directivity pattern of receivers, such as microphones, they are placed on a turntable (Outline ET 250-3D) in the center of the circular loudspeaker array. The impulse response measurement utilizes the multi-sweep method proposed by Majdak et al. [5]. The sweep measurements are repeated for every 10° starting at 0° azimuth until 170° for a full 3D sampling. Prior to the actual measurement, each loudspeaker's transfer function to the center of the array is measured with a reference microphone (NTI MA 2230).

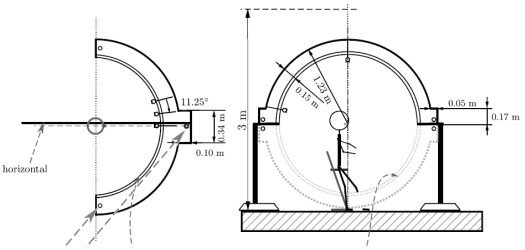


Fig. 1: Double Circle Microphone Array. Schematic of the measurement setup with two wooden circular rings (thickness: 21 mm).

2.2.2 Measurement of Sources

Loudspeakers The measurement of loudspeakers employs the double circular microphone array. As the loudspeakers are placed on a turntable in the center of the array, only the microphones of the vertical ring are used. Typically the acoustical center of a loudspeaker is frequency-dependent, especially in a multi-driver system. Therefore, a frequency-dependent positioning during the measurement or the application of acoustical centering algorithms [3, 6] as a post-processing step might be helpful. For each rotation of the turntable a single logarithmic sweep [7] is played back.

Human Speakers/Singers Impulse responses from directivity measurements for different phonemes used in human speech or singing can be acquired by the glissando method proposed in [8]. The head of the performer is equipped with reflective markers for optical tracking and she/he is asked to sit within the measurement setup close to a centered reference microphone. A visual feedback of his/her position is provided, whereby the mouth is defined as the acoustical center. The performer is asked to sing a glissando starting at a low pitch and ending one octave above the starting pitch. Thus, overlapping of the fundamental frequency and its first harmonic can be avoided. The measurement is made in the horizontal and vertical plane simultaneously to reduce measurement time and positioning drifts.

2.3 Measurement Signal Processing

The impulse responses for the calculation of the directivity patterns are computed by frequency-domain deconvolution of the measured sweep responses with the measurement signal (multi/single sweep, glissando).

A more detailed description of sweep measurement and deconvolution can be found in [7]. In order to exclude room effects such as floor reflections, the impulse responses are truncated to a necessary minimum length using a Hann window. Depending on the application a weighting function should be applied to the impulse response data to compensate for the more dense distribution of the sampling towards zenith and nadir, e.g. for the calculation of the directivity index.

2.4 Visualization using Spherical Harmonics

The visualization uses an interpolation scheme which is applied in the spherical harmonics (SH) domain. Therefore, the sound pressure values are multiplied by an inverse matrix holding SH coefficients $\boldsymbol{Y}_{nm} = [(\boldsymbol{y}_{nm}(\varphi_1, \vartheta_1)\boldsymbol{y}_{nm}(\varphi_2, \vartheta_2)...\boldsymbol{y}_{nm}(\varphi_L \vartheta_L)]^T$ for each distinct measurement point. The inverse transformation of the spherical wave spectra $\boldsymbol{\psi}_{nm}$ in Eq. 1 is done with a denser grid of SH coefficients $\tilde{\boldsymbol{Y}}_{nm}$ to achieve a higher resolution. This process merely corresponds to an interpolation of the pressure values between the spherical measurement points. The vector p describes the frequency-dependent measured sound pressure and the vector p_i the interpolated sound pressure values. The values are evaluated at kr, where k is given by $k = \frac{\omega}{c}$, with k denoting the wave number, c the speed of sound, and r the radius of the measurement aperture, respectively. Further information about the spherical harmonic decomposition can be found in [9].

$$\boldsymbol{\psi}_{nm} = \begin{pmatrix} p(kr, \varphi_1, \vartheta_1) \\ p(kr, \varphi_2, \vartheta_2) \\ \dots \\ p(kr, \varphi_N, \vartheta_N) \end{pmatrix} \boldsymbol{Y}_{nm}^{-1} \tag{1}$$

$$p_i(kr, \varphi, \vartheta) = \boldsymbol{\psi}_{nm}(kr)\tilde{\boldsymbol{Y}}_{nm} \tag{2}$$

3 Evaluation of Directivity Patterns

3.1 Visualization - MATLAB Viewers

The *2D Pattern Viewer* facilitates the comparison of two separate measurements along the horizontal and vertical planes together with features, such as the directivity index (DI) or the correlation coefficient. The directivity index of a sound source is typically defined as the ratio of the on-axis sound power to the overall sound power [2]. Furthermore, the viewer visualizes frequency responses and directivity index over frequency.

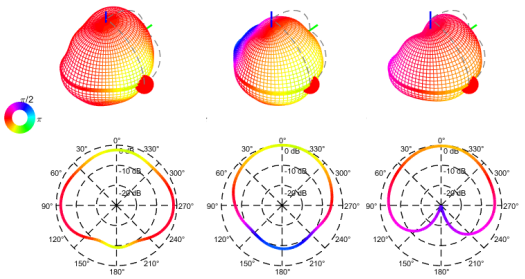
The *3D Pattern Viewer* renders the full 3D patterns in a balloon plot and an adjustable slice plane as 2D polar pattern. The frequency-dependent directivity can be studied by moving the frequency slider.

3.2 Exemplary Measurement Data

The following measurements are all conducted in the anechoic chamber at IEM.

3.2.1 Guitar Cabinets

Fig. 2 shows the measurement results of a self-made guitar cabinet system using a 12-inch speaker driver in a closed-back and open-back configuration. The first one can be presumed to be omnidirectional and the second one to have figure-of-eight characteristic. The measurement results agree with this presumptions for frequencies below 800 Hz. Moreover, the measurement of the stacked cabinets shows that the combination of the two cabinets radiates sound with almost a perfect cardioid pattern. As the two drivers cannot be arranged to be vertically coincident, the cardioid pattern is not achieved rotational symmetric and yields lower sound pressure levels near the zenith and nadir.

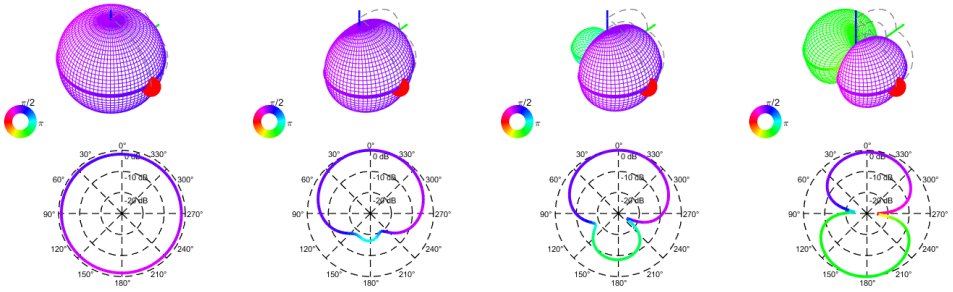


(a) Closed-back only (b) Open-back only (c) Stacked cabinets

Fig. 2: Measured directivity patterns of guitar cabinets for f = 300 Hz.

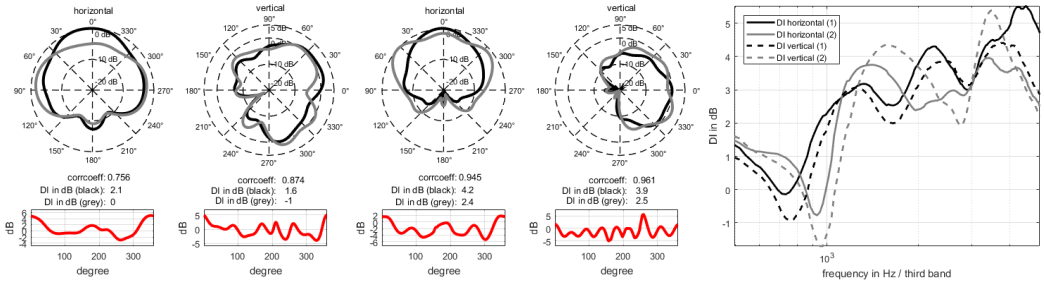
3.2.2 AKG C414 Condenser Microphone

One of the most popular large-diaphragm condenser microphones in recording is the AKG C414 developed in Vienna, Austria. The selectable directivity is one of its key features: the microphone allows switching between omnidirectional, cardioid, supercardioid and figure-of-eight directivity patterns. The measurement results confirm the four directivity pattern settings of the manufacturer, as shown in Fig. 3 for 1 kHz.



(a) Omnidirectional (b) Cardioid (c) Supercardioid (d) Figure-of-eight

Fig. 3: Measured directivity patterns of AKG C414 microphone with selectable directivity at f = 1000 Hz.



(a) Directicity pattern at f = 1000 Hz (b) Directivity pattern at f = 2300 Hz (c) Horizontal and vertical directivity index

Fig. 4: Comparison of two human speakers/singers (back, gray) in the horizontal and vertical plane.

3.2.3 Human Speaker and Singer Directivity

The directivity data for two human speakers/singers is measured with the double circle microphone array using the glissando method, as described above. The polar patterns in Fig. 4 show a difference of at least 4 dB at a frequency of 1 kHz on axis, and at 2.3 kHz in the vicinity of $\pm 50^\circ$. Thus, the differences of the two performers in terms of body shape, posture, and head tilt are noticeable. Alternatively, the radiation characteristics can be compared using the frequency-dependent directivity index. Fig. 4 (c) depicts the horizontal and vertical directivity indeces (HDI, VDI) for the two performers. The repetitive drops along the frequency axis of HDI and VDI are due to destructive interferences by torso reflections. Therefore, the on-axis sound pressure level decreases. For the two performers, these drops of HDI and VDI occur at different frequencies, because of individual physical properties.

Measurements in [10] showed that differences between the directivity patterns of different performers decreased with the measurement distance. Nevertheless, the current results for a measurement distance of 1 m still reveal clear inter-performer differences in the directivity patterns.

4 Conclusion

This report discusses and describes DirPat, a database and viewer of 2D/3D directivity patterns of the IEM. The main idea of this work is to provide data and tools which can be utilized to understand and evaluate directivity characteristics of sound sources and receivers. DirPat is in the spirit of open data and reproducible research: the MATLAB scripts for the 2D/3D viewers, all directivity data discussed here, as well as other measurement data from IEM are freely available here: `https://opendata.iem.at`

2.2 A Voice Directivity Measurement System with Facial Tracking and Augmented Acoustics

This work was published as:

M. Brandner, M. Frank, and A. Sontacchi. A Voice Directivity Measurement System with Facial Tracking and Augmented Acoustics. *Proceedings of the DAGA*, 47:71–74., Vienna. 2021.

The idea and concept of this article were outlined by me, the first author, with help from the second and third author. I wrote the original draft of the manuscript with periodical contribu- tions from the second author. The revision and editing was done by me with help from the third and second author. I did all of the programming, and prepared the data for publication.

DAGA 2021 Wien

A voice directivity measurement system with facial tracking and augmented acoustics

Manuel Brandner[1], Matthias Frank[1], Alois Sontacchi[1]
[1] *Institute of Electronic Music and Acoustics, 8010 Graz, Austria, Email: brandner@iem.at*

Introduction

Voice directivity has an influence on the perceived acoustics for both the singer/speaker and the audience [1, 2, 3]. One of the most important aspects of voice directivity in a room is the direct-to-reverberant energy ratio (D/R ratio) at the listening position. The more focused the voice directivity is, the higher the D/R ratio [4, 5]. This is why voice directivity is becoming more and more important in virtual or augmented acoustic reality [6].

Another field of interest is the performance analysis of professional voices as part of the education. Typical analysis tools for voice characteristics include linear prediction. This method gives incomplete or even erroneous information if the voice is analysed at higher pitch [7, 8]. Voice directivity may help to identify general characteristics in addition to other acoustic features.

Previous studies showed that a change in mouth opening has an effect on the directivity characteristics, i.e., the main direction and the beam width [9, 10]. Different characteristics may hold valuable information about specific vocal tract configurations. Acquiring detailed voice directivity data from human speakers or singers is no easy undertaking due to head movements or mouth changes during measurements. Therefore, we propose an acoustic measurement system enhanced by a facial tracking system to build a ground truth for different mouth opening configurations in speaking and singing. Furthermore, to support natural usage of the voice while actually being in an anechoic measurement room, the system provides augmented acoustics to simulate a more comfortable acoustic environment [6].

This contribution explains the systems, its components, and the methods to process the measurement data. Finally, we present some exemplary results for measurements of different vowels and voice qualities sung by two classical singers.

Measurement system

A measurement system for the determination of sound radiation characteristics of the singing voice was set up by using the double circle microphone array (DCMA) [11], optical tracking sensors in order to measure the oral posture and center position of the singer, and a video camera. The video camera allows to validate the measurement results from the tracking system and opens the possibility to calculate the mouth opening from video directly. The measurement software was implemented in Pure Data [1].

[1] freely available under `http://puredata.info/`

Double circle microphone array

The used microphone array has a radius of 1 m and consists of two circular rings, one placed in the horizontal plane and one in the vertical plane [11]. Each ring can hold up to 32 microphones resulting in an angular spacing of 11.25° and a total number of 62 microphones. In addition, a reference microphone is used, which is located at the exact center of the microphone array if the glissando method is used [9, 12].

Tracking

The mouth opening and absolute position of the singer/speaker inside the measurement array can be captured by a tracking system. The tracking system uses ten cameras, six of them are positioned on the ceiling of the measurement room and four closer in front of the singer/speaker. The closer cameras increase the localization accuracy for the mouth tracking.

Augmented acoustics

For the acoustic analysis, dry signals measured in an anechoic environment are ideal. However, in singing/speaking, room acoustics support the voice, which is a necessity in a longer recording session. Therefore, we use an augmented acoustic system [6] that gives the singer natural room acoustics via transparent headphones [13] without creating reverberation on the microphone signals. The augmented acoustics is fed by the microphone in front of the singer and employs a static, however frequency-dependent directivity [14] to excite the virtual room. The virtual room simulates a shoe-box-like concert hall with a size of roughly 30 m×24 m×20 m and reverberation time of 2.2 s.

Methods

This section presents methods to analyse voice directivity and metrics to discuss directivity characteristics.

Tracking data analysis

The tracking data is used to analyse the steadiness of a sustained sung vowel and to calculate the mouth opening area. Therefore, the tracking data of each time frame of the single facial markers around the mouth are put in order by calculating the convex hull. From the resulting polygon, the area is calculated, which gives the effective mouth opening area over time. This area comes with a bias of around 5 to 7 cm^2 due to the positioning of the markers around the lips, which needs to be considered for further analysis. Furthermore, the height and width of the mouth can be calculated and give valuable information about effects occurring in each single plane (horizontal or vertical). The tracking data during each segment (e.g. sustained vowel) can be used to analyse whether there is an increase or decrease of the mouth area during phonation.

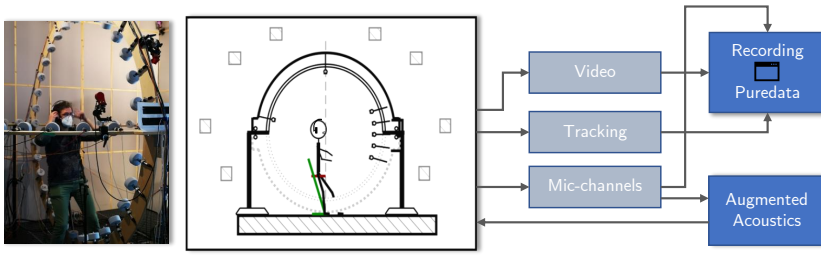


Figure 1: Left: person in the microphone array. Right: schematic representation of the measurement setup.

Calculation of directivity characteristics

Directivity characteristics are computed from frequency data of a sung phoneme at a single pitch calculated by Welch's method (averaged periodogram method)[15] or sung glissandi by the use of the Glissando method. Analysing glissandi can help to overcome the problem of having energetic gaps between harmonics in the frequency domain data of sung sustained vowels at a single pitch [9, 12].

Averaged periodogram method

The microphone signals at the measurement positions are segmented and transformed via the Fourier Transform into the frequency domain. The estimated frequency responses are averaged over time and smoothed afterwards in, e.g., octave, third, or semitone bands. The advantage of the method is that it can be utilized to calculate directivity patterns from shorter speech or sung segments. Nevertheless, it is not possible to easily guarantee that each frequency bin holds valuable information and that it is above the noise floor. Due to the harmonic structure there are gaps between the harmonics that only hold low energy and possibly do not exceed the noise floor. If too much noise energy is averaged the resulting directivity pattern tends to have an omnidirectional shape with a directivity index (DI) of 0 dB. A fundamental frequency tracking algorithm and a proper noise floor calculation can be utilized to partly overcome this problem.

Glissando method

If the Glissando method is used, a reference microphone is needed. The acoustic signals are transformed via the Fourier Transform into the frequency domain. Then, the complex signals are divided by the complex reference signal. This gives the output to input relationship including the phase of the acoustic paths from the reference microphone to each measurement position. The transfer functions transformed back into the time domain give then impulse responses which can be cut to a minimum length to remove room influences, such as floor reflections. This method offers a high signal-to-noise ratio and due to the sung glissando the transfer functions can be calculated for a broad frequency range. Furthermore, this is a much faster method to acquire data in comparison to the averaged periodogram method and less exhausting for a singer/speaker if the glissandi have been trained in advance.

Directivity index for a single plane

The directivity factor $\gamma_p(\omega) = \frac{P_{on-axis}}{P_{mean}}$ in each plane is defined by the ratio of the on-axis power $P_{on-axis}$ to the average power P_{mean} of all sampling positions on the respective plane. The horizontal and vertical directivity index (HDI, VDI), evaluated at an angular frequency ω are defined in dB as follows:

$$DI(\omega) = 10\log_{10}(\gamma_p(\omega)). \tag{1}$$

Energy vector

The energy vector $\boldsymbol{r_E}$ in Eq. (2) can be utilized to describe the direction and the width of the main lobe of an acoustic source. This measure is commonly used in the context of 3D loudspeaker setups, but is as well useful in the description of the characteristics of any arbitrary sound source radiation [16].

$$\boldsymbol{r_E} = \frac{\sum_{i=1}^{L} |H(\omega,\phi_i)|^2 \boldsymbol{m_i}}{\sum_{i=1}^{L} |H(\omega,\phi_i)|^2} \tag{2}$$

The frequency-dependent magnitudes $H(\omega,\phi_i)$ are multiplied by the vectors $\boldsymbol{m_i} = [\cos(\phi_i), \sin(\phi_i)]^T$ of each measurement position i, $i = 1, 2, ..., L$ in each respective plane, and normalized by the sum of the energy, yielding a normalization of the vector between the limits 0 (omnidirectional) to 1 (maximum focus to one direction). The following two metrics are used, here: (i) the main beam width in each plane $\theta_w = 2\arccos\|\boldsymbol{r_E}\|$ and (ii) the main direction in the vertical plane $\theta_s = \arctan y_{\boldsymbol{r_E}}/x_{\boldsymbol{r_E}}$.

Exemplary Results

In this section we present exemplary results from two classical singers for the methods described in the section above.

Analysis of tracking data and audio signals

In Fig. 2 we show the representation of the time domain audio signal and the voice activity detection (VAD) of sung segments during one recording run. The vowel sequence /a:/, /e:/, /i:/, /o:/, and /u:/ is repeated three times as each sequence is sung with a different voice quality (modal, breathy, and pressed). Furthermore, we show in Fig. 2 the corresponding mouth area calculated from the data stream of the tracking markers around the mouth. During a sustained sung vowel a change of mouth opening can occur, which is indicated by a regression line above each segment. If the mouth opening increases or decreases the straight line is tilted either way.

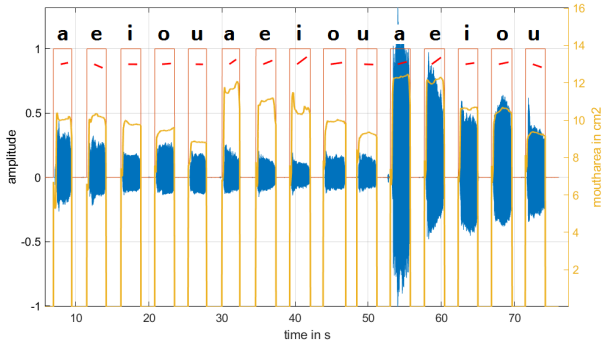


Figure 2: Time domain representation of the audio signal, VAD, and mouth area tracking. Tilted lines over each segment indicate a dynamic change during phonation - mouth area has been increased or decreased by the singer.

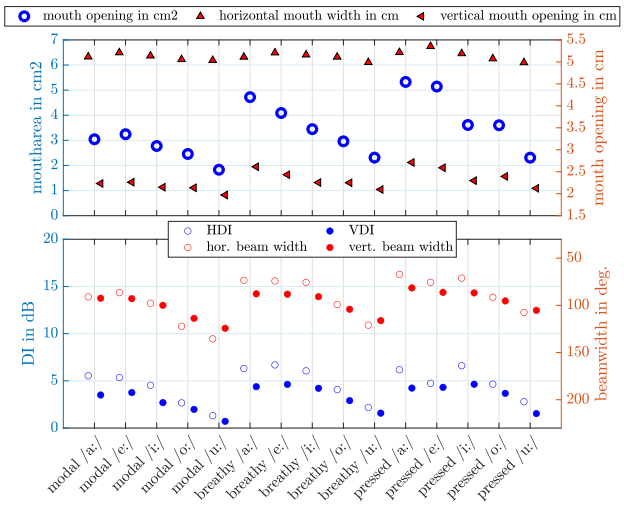


Figure 3: Top: averaged mouth opening area for each segment and the corresponding mouth width and height in cm. Bottom: horizontal, vertical, and averaged directivity index and for the corresponding plane the beamwidth of the energy vector and its average calculated from one-third-octave smoothed data.

Directivity analysis

For a change of vowel or mouth opening we expect a change in directivity or in other words some effect on the directivity characteristics (e.g. directivity index, beam width, beam direction).

Averaged periodogram method

The directivity analysis with the averaged periodogram method of vowels sung by a classical singer at the pitch a/A3 of approx. 220 Hz is shown in Fig. 3. Again, the vowel sequence /a:/, /e:/, /i:/, /o:/, and /u:/ sung with three different voice qualities is investigated. In the upper plot of Fig. 3 a decrease of the effective mouth area used by the singer can be seen. Furthermore, a slight decrease of the mouth width and more pronounced decrease of the mouth height can be recognized. In the lower plot of Fig. 3 a decrease of the directivity index and the beamwidth in both planes from vowel /a:/ to /u:/ is visible for all three voice qualities.

Glissando method

A soprano singer was asked to sing the german vowel /a:/ with different provoked mouth openings (similar as in [9]). This is of special interest as the classical singing technique is often associated with a lowered jaw [17] and much larger mouth openings are expected to occur in singing than in conversational speech. In the horizontal plane a larger mouth width provokes a decrease in the beam width starting at around 2 kHz, whereas larger mouth openings along the vertical axis only show a change in beam width starting at around 5 kHz (see Fig 4(a) and (b) - top plots). In the bottom plots in Fig 4(a) and (b) the corresponding main direction of the found beam widths are shown. In the horizontal plane symmetric patterns provoke a focusing of the radiated sound towards the 0° direction for most frequencies. This differs in the vertical plane as frequencies below 1 kHz and above 5 kHz are more focused towards the floor, whereas between 1 kHz and 5 kHz sound is more radiated upwards or to the front. In Fig. 5(a) we show polar patterns normalized to the maximum with differences of around 5.4 dB towards the sides and up to 8.4 dB in Fig. 5(b) in the vertical plane for the forward upwards direction. In Tab. 1 we list the additional information from the tracking system which provides insight on the quality of the measurement results. While the effective mouth opening areas differ between the provoked mouth openings, the head inclination angle changes no more than 4.1°.

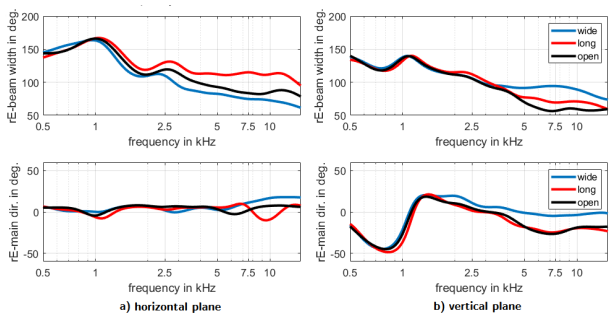


Figure 4: Glissandi analysis. In the top plots (a) horizontal plane and (b) vertical plane the differences between provoked mouth openings are displayed. In the bottom plots we provide the corresponding main direction of the displayed beam widths in the top plots.

Conclusion

We presented an enhanced measurement system to measure voice directivity and discussed strategies and methods to investigate directivity characteristics in detail by exemplary measurement results from two classical singers. The results show that differences between vowels and between different mouth openings can be made visible and substantiated by the use of a tracking system to verify the center position and validate the effective mouth

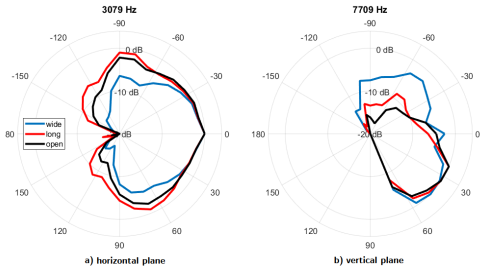


Figure 5: Polar patterns for the three mouth openings (a) in the horizontal plane at 3 kHz and (b) in the vertical plane at 7.7 kHz.

Tracking \ mouth	wide	long	open
Δ Area in cm^2	+0.52	+5.06	+10.66
Δ Head inclination in °	+0.40	+1.79	+4.10

Table 1: Mouth area and head inclination of the three investigated mouth openings listed as delta values. To calculate the delta values we use as a reference the mouth area for the vowel /a:/ used in speech by the singer.

opening area. Furthermore, the tracking data allows us to analyse the steadiness of a singer during recordings. We also enhanced the measurement setup by including augmented acoustics via transparent headphones, which provides natural room acoustics during the recording session. This opens up the possibility to investigate the influence of different room acoustics on the singing strategy of a singer, for example, on the vowel intelligibility. The microphone array offers to encode the actual directivity pattern of the singer into the augmented acoustics which is a goal in the future.

Acknowledgments

Special thanks to Thomas Musil for his help in implementing the measurement software in Pure Data.

References

[1] Schärer Kalkandjiev, Z. and Weinzierl, S.: The Influence of Room Acoustics on Solo Music Performance: An Empirical Case Study. Acta Acustica united with Acustica 99/3 (2013), 433–441

[2] Fischinger, T., Frieler, K., and Jukka L.: Influence of virtual room acoustics on choir singing. Psychomusicology: Music, Mind, and Brain 25/3 (2015), 208-218

[3] Postma, B., Demontis, H., and Katz, B.: Subjective Evaluation of Dynamic Voice Directivity for Auralizations. Acta Acustica united with Acustica 103 (2017), 181-184

[4] Frank, M. and Brandner, M.: Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions. 5th International Conference on Spatial Audio (2019), 131-135

[5] Frank, M. and Brandner, M.: Perceptual Evaluation of Spatial Resolution in Directivity Patterns. Fortschritte der Akustik - DAGA (2019)

[6] Frank, M., Rudrich, D., and Brandner, M.: Augmented Practice-Room - augmented acoustics in music education. Fortschritte der Akustik - DAGA, 2020

[7] Arroabarren, I. and Carlosena, A.: Inverse filtering in singing voice: a critical analysis. IEEE Transactions on Audio, Speech, and Language Processing, vol. 14 (2006), 1422-1431, https://doi.org/10.1109/TSA.2005.858013

[8] Bereuter, P., Kraxberger, F., Brandner, M., and Sontacchi, A.: Design of a vowel and voice quality indication tool based on synthesized vocal signals. Journal of the Audio Engineering Society (2021)

[9] Brandner, M., Blandin, R., Frank, M., and Sontacchi, A.: A pilot study on the influence of mouth configuration and torso on singing voice directivity. The Journal of the Acoustical Society of America 148/3 (2020), 1169-1180

[10] Pörschmann, C. and Ahrend, J.: Investigating phoneme-dependencies of spherical voice directivity patterns. The Journal of the Acoustical Society of America 149/6 (2021), 4553-4564. Available: https://doi.org/10.1121/10.0005401

[11] Brandner, M, Frank, M., and Rudrich, D.: DirPat-Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers. Audio Engineering Society Convention 144,(2018)

[12] Malte, K. and Jers, H.: Directivity measurement of a singer. The Journal of the Acoustical Society of America 105/2 (1999), 1003-1003

[13] Meyer-Kahlen, N., Rudrich, D., Brandner, M., Wirler, S., Windtner, S., and Frank, M.: DIY Modifications for Acoustically Transparent Headphones. AES 148th Convention, e-Brief 61 (2020)

[14] Weinzierl, S., Vorländer, M., Behler, G., Brinkmann, F., von Coler, H., Detzner, E., Krämer, J., Lindau, A., Pollow, M., Schulz, F., et al.: A database of anechoic microphone array measurements of musical instruments, 2017, https://doi.org/10.14279/depositonce-5861.2

[15] Welch P. D.: The use of fast Fourier transforms for the estimation of power spectra: A method based on time averaging over short modified periodograms. IEEE Transactions on Audio and Electroacoustics, vol. 15 (1967), 70-73

[16] Gerzon, M. A.: General metatheory of auditory localisation. Audio Engineering Society Convention 92, (1992)

[17] Nair, A., Nair, G. and Reishofer, G.: The Low Mandible Maneuver and Its Resonential Implications for Elite Singers. Journal of Voice, vol. 30/1 (2016), 128.e13–128.e32. Available: http://dx.doi.org/10.1016/j.jvoice.2015.03.010

2.3 Real-Time Calculation of Frequency-Dependent Directivity Indexes in Singing

This work was published as:

M. Brandner, A. Sontacchi, and M. Frank. Real-Time Calculation of Frequency-Dependent Directivity Indexes in Singing. *Proceedings of the DAGA*, 45:70–73., Rostock. 2018.

The idea and concept of this article were outlined by me, the first author, with help from the second author. I wrote the original draft of the manuscript with periodical contributions from the third author. The revision and editing was done by me with help from the third and second author. I did all of the programming, and prepared the data for publication.

DAGA 2019 Rostock

Real-Time Calculation of Frequency-Dependent Directivity Indexes in Singing

Manuel Brandner, Alois Sontacchi, and Matthias Frank
Institute of Electronic Music and Acoustics, 8010 Graz, Austria, Email: brandner@iem.at

Introduction

Directivity of speech and singing is of great interest in acoustic modeling [1], performance studies [2, 3] and for architectural design [4]. Voice directivity is frequency-dependent and primarily determined by the physiology of a person. Furthermore, it can be changed to an extend by posture, head inclination, and vocal tract configuration. The most prominent influences on voice directivity can be summarized as:

- shape and size of head and torso (fixed),
- posture, head inclination (changeable),
- vocal tract geometry (changeable),
- spectral emphasis (changeable).

While, a recent database [5] includes the frequency-dependent directivity for each playable tone of some instruments and some of its overtones in order to facilitate maximum authenticity in reproduction, typical models in acoustic simulations do not take the aforementioned effects and properties of speech and singing into account. These models are defined by one global directivity pattern for a large frequency range.

In our study we investigate different methods for calculating the voice directivity of a classical singer. We show the applicability of the short-term Fourier transform analysis (STFT analysis) analyzing singing at steady pitch in comparison to a less time-consuming alternative, namely the glissando method proposed in [6]. We define metrics to evaluate our results and to better investigate the variability of voice directivity over frequency by the example of a classical singer. We use the directivity index for the horizontal and vertical plane (HDI, VDI) as objective metrics [11]. For the real-time calculation, the HDI and VDI values are evaluated at the fundamental frequency and its harmonics to gain more robust information with regard to the SNR (signal-to-noise ratio). The analyzed short-term data is then compared to results calculated by the glissando method.

One professional classical singer with a master's degree in classical voice and international singing experience was asked to participate in our study. We asked the singer to use two different strategies to vocalize the German vowel /a/. One strategy is using a rather small mouth opening and the other a rather large mouth opening achieved by lowering the jaw as discussed in [7]. To analyze the outcome of the strategies we compare the results of each strategy for each measurement method.

Method

Measurement Setup

The measurement of source radiation patterns employed a microphone array consisting of two circular rings, one placed in the horizontal, the other one in the vertical plane, respectively, cf. Fig. 1 and [8]. Each of the rings can hold up to 32 microphones resulting in a maximum number of 62 microphones as both rings intersect in the front and back of the array. The angular spacing of the microphones is 11.25°. To reduce reflections the center-facing side of the rings with a thickness of 21 mm are beveled. The distance from the microphone capsules to the center-facing side is 8 cm to ensure a radius of 1 m from the capsules to the center of the array. The full diameter of the apparatus spans 2.56 m at its widest point. The center of the array can be lifted to any height between 1.3 m and 2 m by adjustable stands. As microphones, we used NTI MA 2230 connected to an Andiamo.MC Directout Technologies microphone preamplifier. An optical tracking system (optitrack flex 13, 6 cameras) was used to validate the position of the head during the recordings to account for possible head movements. Furthermore, the system allows to some extent to guide the test subject by a visual feedback of the current head and actual center position during measurements. In the case of a singer, the mouth is defined as the acoustical center.

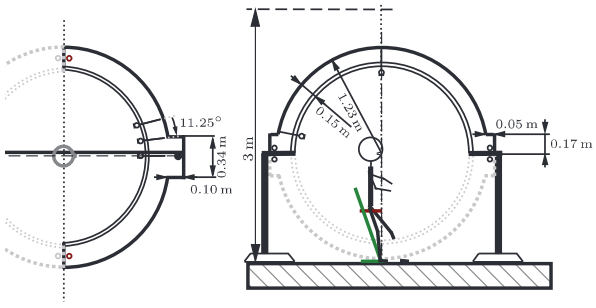


Figure 1: Double Circle Microphone Array. Schematic of the measurement setup with two wooden circular rings.

Short-term Fourier transform analysis

The singer is asked to sustain the German vowel /a/ for as long as possible at steady pitch (note G4) and to attempt to use similar vocal effort for all measurement runs. The microphone signals are analyzed in blocks of equal length and transformed from time domain to the frequency domain. The magnitude spectra are averaged over time for each channel and then third-octave smoothed.

The radiation characteristics can be compared using the frequency-dependent directivity index. The calculation is done in quasi real-time (averaging only over a limited number of blocks) with the 2D Polar Pattern and Spectrum Analyzer [8] where the number of Fourier coefficients and the hopsize is adjustable. The examplary data reported in this document is calculated with a window and FFT size (Fast Fourier transform) of 2048 samples at a sampling frequency of 44100 Hz. We use a 4-term Blackman-Harris window and therefore an overlap of 66.1% between consecutive frames [9]. The power spectra $|H(\omega, n)|^2$, where n denotes the discrete time instants, are averaged by using a first-order recursive filter (Eq. 1). The filter coefficient λ, which can be seen as a forgetting factor, is set to 0.5 (smoothing time $\approx$ 100 ms).

$$\hat{H}(\omega, n) = \lambda |H(\omega, n)|^2 + (1 - \lambda)|H(\omega, n - 1)|^2 \qquad (1)$$

The harmonic structure of the spectrum in singing does not provide valuable information (amplitude above the noise floor) at each frequency bin and at all time. There are gaps between the overtones that only hold low energy and possibly fall below the noise floor from time to time. If too much energy of the noise floor is averaged, the directivity changes towards an omni-directional pattern in the analysis. This happens as measurement noise provokes a directivity index of around 0 dB. To overcome this misleading effect a pitch tracker is used to calculate the HDI and VDI solely at the fundamental frequency and its harmonics.

Glissando method

The glissando method (vocal sweep method, [6]) allows to calculate impulse responses directly from directivity measurements for vocalized phonemes. To capture the source signal a reference microphone is positioned in front of the singer as close as possible ($<$ 3 cm). The performer is asked to sing a glissando (vocal sweep) starting at a low pitch (G4) and ending at a higher pitch at least one octave above. The impulse responses are calculated by deconvolution of the measured signals at each microphone by the reference signal in the frequency domain and then cut to a length of 512 samples and windowed in the time domain. This simultaneous measurements allow a reduction of both measurement time and positioning errors.

Visualization using circular harmonics

The visualization of the polar patterns uses an interpolation scheme which is applied in the circular harmonics domain (cf. [8]). Further information about circular and spherical harmonic decomposition can be found in [10].

Horizontal and vertical directivity index

The directivity factor $\gamma_p(\omega)$ is calculated at each frequency for the horizontal and vertical plane, respectively. It is defined by the ratio of the on-axis power to the mean power of all sampling positions, where L denotes the total number of measurement points within the corresponding plane.

For both planes we denote the angle ϕ_i for each measurement position independent of the orientation of the plane with ϕ_0 as the on-axis direction.

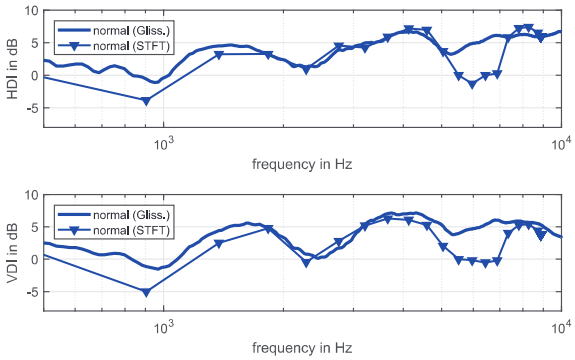
$$\gamma_p(\omega) = \frac{|H(\omega, \phi_0)|^2}{\frac{1}{L}\sum_{i=0}^{L-1}|H(\omega, \phi_i)|^2} \qquad (2)$$

The directivity index at a frequency ω for each plane is then defined in dB as follows

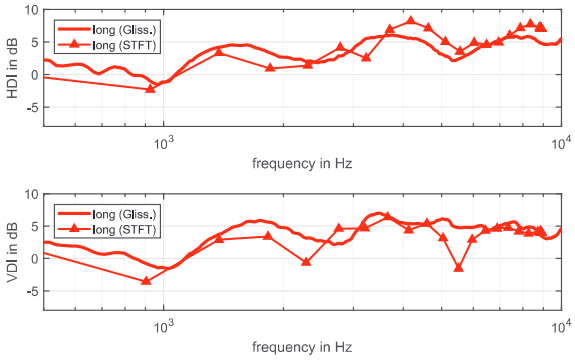
$$DI(\omega) = 10 \log_{10}(\gamma_p(\omega)). \qquad (3)$$

Results

Within this section we present a comparison of the results for the two proposed methods to calculate the frequency-dependent directivity index. For each method the frequency data used for calculating the directivity indexes is third-octave smoothed. The classical singer was asked to use two different strategies to produce a similar German vowel /a/. Our suggested strategies are to use a rather small mouth opening "normal", more similar to speech, and in comparison a vertical larger mouth opening "long" achieved by lowering the jaw.



(a) Comparison of HDI/VDI values for "normal" condition



(b) Comparison of HDI/VDI values for "long" condition

Figure 2: Comparison of HDI/VDI values calculated with the STFT analysis (only evaluated at $\triangle$ or $\triangledown$) and glissando method for two strategies for the vowel /a/. The results of both methods show good agreement, besides minor deviations.

Comparison of the proposed methods

In Fig. 2 we compare the results of the glissando method with the results of the STFT analysis. In most of the frequency regions the results agree very well. Although, the HDI/VDI values differ for the "normal" mouth opening at frequencies below 2 kHz, as seen in Fig. 2a. It is especially shown that around 1 kHz larger differences occur between the methods. This can be explained by the different frequency resolutions used for the two measurement methods and the spectral sparsity of a sung note (STFT). The dip around 1 kHz, which is provoked by the shadowing and reflection properties of the torso, occurs within a bandwidth of around 200 Hz. As the metrics are calculated from data smoothed over third octaves and the methods use different frequency resolutions this dip is less pronounced for the glissando method (Δf=86 Hz) than for the STFT analysis (Δf=21.5 Hz). In Fig. 3 we show the influence of the smoothing bandwidth on the results of the glissando method for the "normal" mouth opening. The decrease of the directivity index around 1 kHz, as seen in the results by STFT analysis, is can be made visible if the smoothing bandwidth is reduced to sixth-octave bandwidth.

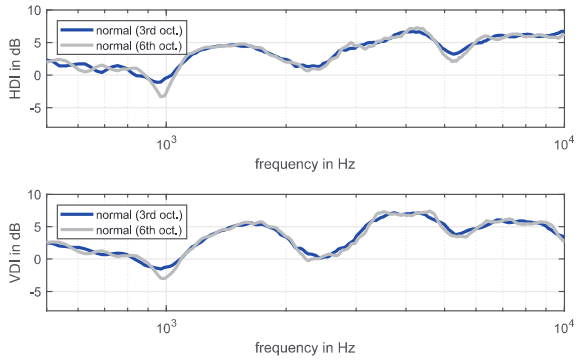


Figure 3: Comparison of HDI/VDI values for the glissando method with third-octave and sixth-octave band smoothing (light gray) for "normal" condition.

For the larger mouth opening "long" much more deviations between the methods are visible (Fig. 2b). These deviations can occur because the exact same mouth opening cannot be guaranteed for both runs and due to differences resulting from averaging; and again due to the spectral sparsity of a sung note. Furthermore, for larger mouth openings higher-order modes are more likely to be radiated from the mouth in comparison to small mouth openings [1].

At higher frequencies the largest deviations are shown around 5 to 7 kHz for the "normal" condition and at 5.5 kHz for the "long" condition which occur due to a low SNR (cf. Fig. 4). Therefore, a threshold for low SNR should be used in the future to easily identify the validity of the measurement data.

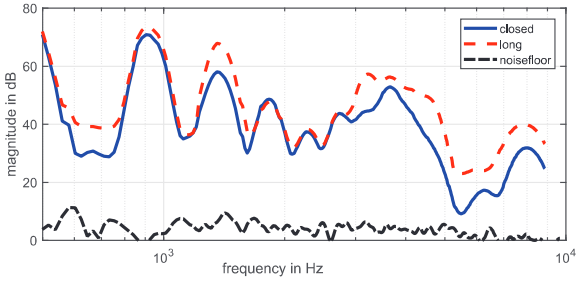


Figure 4: Comparison of the on-axis response third-octave smoothed) for the two strategies for the vowel /a/ from STFT analysis.

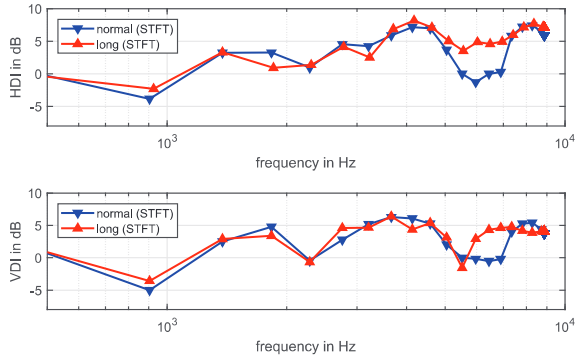


Figure 5: Comparison of HDI/VDI values for the two strategies for the vowel /a/ with the STFT analysis (from third-octave smoothed data).

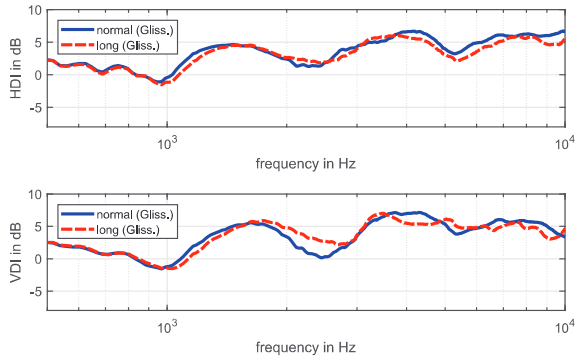


Figure 6: Comparison of HDI/VDI values for two mouth openings for the vowel /a/ with the glissando method.

Comparison of the singing strategies

This section compares the results of each method with regard to the singing strategy (mouth opening) used. The largest deviations are about 2 dB and can be seen at several frequencies, if we exclude the frequency region of the STFT analysis where we attested bad SNR. Nevertheless, the explained better SNR for the "long" condition within the frequency region of 5 to 7 kHz implies higher spectral energy in comparison to the "normal" singing condition. Therefore, the "long" condition yields higher directivity if we incorporate the level in our analysis, see Fig. 5.

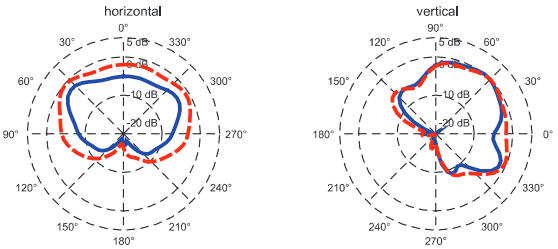


Figure 7: Comparison of the polar patterns at 2.5 kHz for two mouth openings: (i) "normal" (solid) and (ii) "long" (dash-dotted) for the vowel /a/ calculated with the glissando method (third-octave smoothed).

As the glissando method is related to a classic sweep measurement it inherits a better SNR than an averaging method like the STFT analysis. In Fig. 6 we see larger deviations in the vertical plane for the results of the glissando method with an increase of directivity especially for the frequency region around 2.5 kHz. In Fig. 7 the differences at 2.5 kHz are shown in polar plots, which are normalized to the maximum of both patterns in each plane. This rather large variation has been also reported in [3] and seems to be linked to the size of the mouth.

If we consider the radiation of a small piston in an infinite baffle as a simplified model, we expect for the change of mouth opening from "normal" to "long": (i) a slight decrease of directivity in the horizontal plane due to a decrease of the mouth width and (ii) an increase of directivity in the vertical plane as the height is increased because of lowering the jaw.

However, this cannot be attested from our data, as differences do not occur intuitively when using the common metrics (HDI/VDI). If we study the radiation characteristics more thoroughly from frequency versus angle representations, we can see that for the larger mouth opening more energy is radiated towards the floor as the frequency increases. This cannot be represented by the classical directivity index.

Conclusion

We show that similar results can be achieved if the directivity index is calculated from short-time averages in comparison to the glissando method or long-term averages (cf. [2]). Although, the analysis needs to consider the sparsity of the spectrum (low frequencies) and the noise floor (high frequencies) which are the reason for larger deviations between the investigated methods. The real-time calculation is especially useful for a fast view on the data already during or after the measurement. Furthermore, it is a practical way when investigating a large number of conditions, e.g. several different phonemes.

Our results show that the directivity index in the horizontal and vertical plane is increasing with frequency except for some drops around 1 kHz, 2.5 kHz, and 6 kHz. We show subtle variations in the directivity index dependent on the used singing strategy (mouth opening). Our results are in good agreement with results found in literature. Future studies should evaluate the directivity of different phonemes with other metrics than the HDI/VDI and investigate the qualitative performance of the impulse responses achieved by the glissando method in auralization.

References

[1] Blandin, R.: Influence of Higher Order Acoustical Propagation Modes on Variable Section Waveguide Directivity : Application to Vowel [A]. Acta Acoustica united with Acustica 102 (2016) 1-12.

[2] Cabrera, D.: Vocal Directivity Measurements of Eight Opera Singers, Proceedings of the 18th International Congress on Acoustics (2004) 505-506.

[3] Katz, B., and D'Alessandro, C.: Directivity Measurements of the Singing Voice, 19th International Congress on Acoustics (2007) 45-50.

[4] Chu, W. T. and Warnock, A. C.: Detailed Directivity of Sound Fields Around Human Talkers, Technical Report for the Institute for Research in Construction (National Research Council of Canada) (2002) 1–47.

[5] Weinzierl, S., et al.: A Database of Anechoic Microphone Array Measurements of Musical Instruments (2017), URL: `http://dx.doi.org/10.14279/depositonce-5861.2`

[6] Kob, M. and Jers, H.: Directivity measurement of a singer. Collected Papers from the Joint Meeting "Berlin 99" (1999), ISBN:3-9804568-5-4

[7] Nair, A. and Nair, G., and Reishofer, G.: The Low Mandible Maneuver and Its Resonential Implications for Elite Singers, Journal of Voice 30 (2016) 128.e13-128.e32.

[8] Brandner, M., Frank, M., and Rudrich, D.: DirPat - Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers. Audio Engineering Society Convention 144 (2018), URL: `http://www.aes.org/e-lib/browse.cfm?elib=19538`

[9] Heinzel, G., Ruediger, A., and Schilling, R.: Spectrum and Spectral Density Estimation by the Discrete Fourier transform (DFT), including a comprehensive List of Window Functions and some new Flat-Top Windows, URL: `http://edoc.mpg.de/display.epl?mode=doc&id=395068`

[10] Zotter, F.: Analysis and Synthesis of Sound-Radiation with Spherical Arrays. University of Music and Performing Arts, Graz (2009), URL: `https://iem.kug.ac.at/fileadmin/media/iem/projects/2010/zotter.pdf`

[11] Tylka, J., Sridhar, R., and Choueiri, E.: A Database of Loudspeaker Polar Radiation Measurements. Audio Engineering Society Convention 139 (2015), URL: `http://www.aes.org/e-lib/browse.cfm?elib=17906`

2.4 Influence of the Vocal Tract on Voice Directivity

This work was published as:

R. Blandin, **M. Brandner**. Influence of the Vocal Tract on Voice Directivity. *Proceedings of the 23rd International Congress on Acoustics*, 23:1795–1801., Aachen. 2019.

The idea and concept of this article were outlined by the first author, with help from me, the second author. I wrote periodical contributions and have conducted the measurements. The revision and editing was done by the first author with help by me.

PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Influence of the vocal tract on voice directivity

Rémi Blandin[(1)], Manuel Brandner[(2)]

[(1)]Institute of Acoustics and Speech Communication, TU Dresden, 01062 Dresden, Germany, remi.blandin@tu-dresden.de

[(2)]University of Music and Performing Arts, Institute of Electronic Music and Acoustics, Graz, brandner@iem.at

Abstract

Voice directivity induces variations of the sound amplitude and frequency content with the direction. Voice directivity is important for the efficient transmission of speech and singing. Therefore, it is taken into account by concert hall designers in order to enhance the quality of artistic performances. While in voice directivity studies the shape of the head and the torso have long been considered to study this phenomenon, the influence of the vocal tract shape/configuration has received very little attention. However, it has been recently shown that the vocal tract configuration influences voice directivity through the internal acoustic field and the frequency content. Within this paper, the contribution of the vocal tract to the voice directivity is characterized through physical modelling and direct measurements on a professional classical singer. In particular, the role of higher order propagation modes is discussed as well as changes of voice directivity due to different vocal tract configurations.

Keywords: Voice, directivity, singing, higher order modes

1 INTRODUCTION

The directivity of speech and singing induces variations of the amplitude and the frequency content of the radiated sound with the direction. Various measurements performed with microphone arrays on speakers and singers show that this phenomenon increases toward high frequencies [1, 2, 3, 5].

It was also shown that the directivity patterns of the various vowels and consonants are different [2, 4, 5, 6]. Also, analyzing data within smaller frequency intervals, such as third of octave bands and even linear frequency discretization [6] highlighted more complex changes of the directivity patterns with the frequency. Moreover, it was shown that the internal acoustic field can potentially influence the directivity of the radiated sound above 4 kHz to 5 kHz, inducing significant changes of directivity patterns within frequency intervals of the order of 100 Hz [7].

Katz and d'Alessandro highlighted differences in directivity patterns measured with different vocal techniques [4]. A global increase of directivity observed with the projected singing technique was attributed to an increase of acoustic energy in the 2.5 kHz third of octave band. However, to our knowledge, no study has focused on the potential influence of the mouth configuration on the directivity. Little speech studies compared measurements with theoretical radiation models, and only with simple models describing the mouth as a vibrating piston.

Thus, the objective of this work is to investigate the influence of the size of the mouth opening on voice directivity by comparing measurements and simulation. The directivity of a professional singer singing the vowel /a/ with two different mouth configurations was recorded with a microphone array. These measurements are compared with two theoretical models:

- a simple vibrating plane piston model set inside an infinite baffle [8],
- and the multimodal theory which allows one to take into account the potential influence of the vocal tract as well as the radiation patterns of the higher order modes [16].

The diffraction by the head and the torso is currently not taken into account in order to focus only on the potential effects of the mouth shape, whereas stronger influences are expected to occur only up to 2 kHz to 2.5 kHz for an average person (head and torso dimensions).

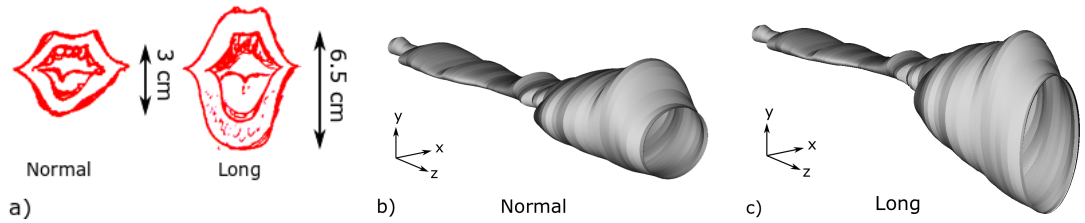


Figure 1. a) Normal and long mouth configurations used by a singer with vertical lower lip to upper teeth distance. b) and c), vocal tract geometries from magnetic resonance image [14] modified to simulate the b) normal and c) long mouth configurations, respectively. The lower part of the oral cavity is rotated to match the lower lip to upper teeth distance measured on the singer.

2 METHOD

2.1 Directivity measurement

The sound radiated by a professional female singer was recorded with a double circle microphone array constituted of 62 microphones regularly distributed on two circles in the horizontal and vertical planes [9]. The singer was asked to sing the vowel /a/ doing a glissando with a normal mouth opening and a long mouth opening. The long mouth opening corresponds to a lower position of the jaw which increases the vertical mouth opening compared to the normal configuration (see Fig 1a). Lowering the jaw is a common practice in classical singing [10]. The mouth opening was characterized with the vertical lower lip to upper teeth distance. This represents the effective mouth opening since the upper lip is above the upper teeth and the lower lip covers the lower teeth. The normal mouth opening corresponds to a vertical distance of about 3 cm and the long one to about 6 cm (see Fig. 1a). The transfer functions between each point of the microphone array and a reference microphone placed as close as possible (< 3 cm) to the singer were computed using a glissando method [11, 12].

2.2 Simulations

An initial geometry obtained by simplifying a three-dimensional geometry extracted from magnetic resonance image [13] was transformed to approximate the different vocal tract configurations investigated. This initial geometry corresponds to the vowel /a/ pronounced by a male speaker. It is constituted of a succession of straight tubes with elliptical cross-sections [14].

Since the singer involved in this study was a female, the length and the volume of the oral and pharyngeal cavity of the original geometry were adapted to female dimensions. The ratios of the average dimensions of 120 male and female subjects [15] were used to adapt the length and cross-sectional area of the sections to female average dimensions.

The initial geometry is axis-symmetric: the centers of each section share a common axis. However, it has been shown that this configuration is not well suited to reproduce the acoustic properties of the vocal tract above 4-5 kHz [16, 7]. Thus, 25 % and 75 % of eccentricity have been introduced in the horizontal (x, z) and the vertical (y, z) plane, respectively. The centers of the cross-sections have been shifted toward the positive values of x and y by 25 % and 75 % of the half width and the half height of the ellipses, respectively.

In order to adapt the mouth opening to the vertical lower lip to upper teeth distance measured on the test subject, a downward rotation of the lower points of the cross-sections of the oral cavity has been operated. The center of rotation was placed at the limit between the pharyngeal and the oral cavities. The height of the ellipses was increased to match the position of the rotated points (see Figs. 1b and 1c).

The transfer functions between an input volume velocity imposed at the glottis and the acoustic pressure radiated by the mouth was computed with the multimodal method described by Blandin et al. [16]. In order to simulate

the vibrating piston model, the same method was used with the plane mode only.

2.3 Directivity maps

The transfer functions computed from the measurements and the simulations are presented as directivity maps in Fig. 2. These maps present the radiated amplitude in color-scale as a function of the angular position and the frequency. Frequency resolutions of 43.1 Hz and 10 Hz are used for the analysis of the measurements and simulations, respectively. The amplitudes are normalized over all angular positions at each frequency by the maximum of amplitude over the different angular positions. Since the simulations use and infinite baffle boundary condition, it can provide data only between -90°and 90°(0°corresponds to the direction normal to the frontal plane). Therefore, the visualization of the measurements will also be limited to the same angular region.

3 RESULTS

Within this section the results of the measurements and the simulations of the two proposed methods are presented.

All the directivity maps presented in Fig. 2 show a beam which becomes narrower as the frequency increases for the measurement and both simulation types. The directivity maps of the measurements and both simulations in Fig.2 show an increasing main beam width as the frequency increases. This main beam is globally narrower for the long configuration (Figs. 2d to 2f).

The measurements (Figs. 2a and 2d) and the simulation performed with higher order modes (Figs. 2b and 2e) show more complex patterns than the simulation obtained with the vibrating piston model (Figs. 2c and 2f). Abrupt changes of the direction (abrupt upward or downward shifts of the beam) and the shape of the directivity patterns as well as localized minima occur in small intervals of the order of 100 Hz. Similar trends can be seen both in the measurement and the multimodal simulations. In particular, in the normal configuration, a localized minimum around 9.5 kHz and -30°, exists both in the measurements and the multimodal simulation (Figs. 1a and 1b). In the long configuration, similar trends can be seen around 3.5 kHz or between 5-6 kHz in the 45°to 90°angular region. However, these complex variations appear only at relatively high frequency in the multimodal simulations: from 5 kHz for the normal configuration (see Fig. 2b), and from 3-4 kHz for the long configuration (see Fig. 2e). The measurements show more complex radiation patterns than the simulations between 0 and 3-4 kHz.

Between 0 and 2 kHz the measurements performed in both mouth configurations are very similar. They have the same radiation patterns, including minima and maxima evolving downward with increasing frequency. These patterns are not seen on the simulations which show little amplitude variations in this frequency range. Small deviations between the measurements performed with the normal and the long configurations occur between 2-3 kHz within the 0°to -90°angular region. More substantial differences are visible from 3kHz on.

In the measurements, from 2 kHz on, globally more energy is radiated downward (between 0°and -90°) in the long configuration compared to the normal one. This is especially seen in the frequency region around 4 kHz. However, with increasing frequency the main beam direction fluctuates between being radiated upward and downward. This trend is seen only for the experimental data and the simulations performed with higher order modes in some frequency bands. Thus, a downward orientation of the beam is present between 3-4 kHz, 5-6 kHz, 7-8 kHz and 9-10 kHz for the multimodal simulations of the long configuration (see Fig. 2e). However, the beam also tend to be orientated upward between 4-5 kHz, 6-7 kHz and 8-9 kHz.

4 DISCUSSION

Within this section the results obtained are discussed. Not only the proposed models are compared with each other and the measurements, but also the differences that occur if the jaw is lowered and the distance between lower lip and upper incisors increases by more than double the height.

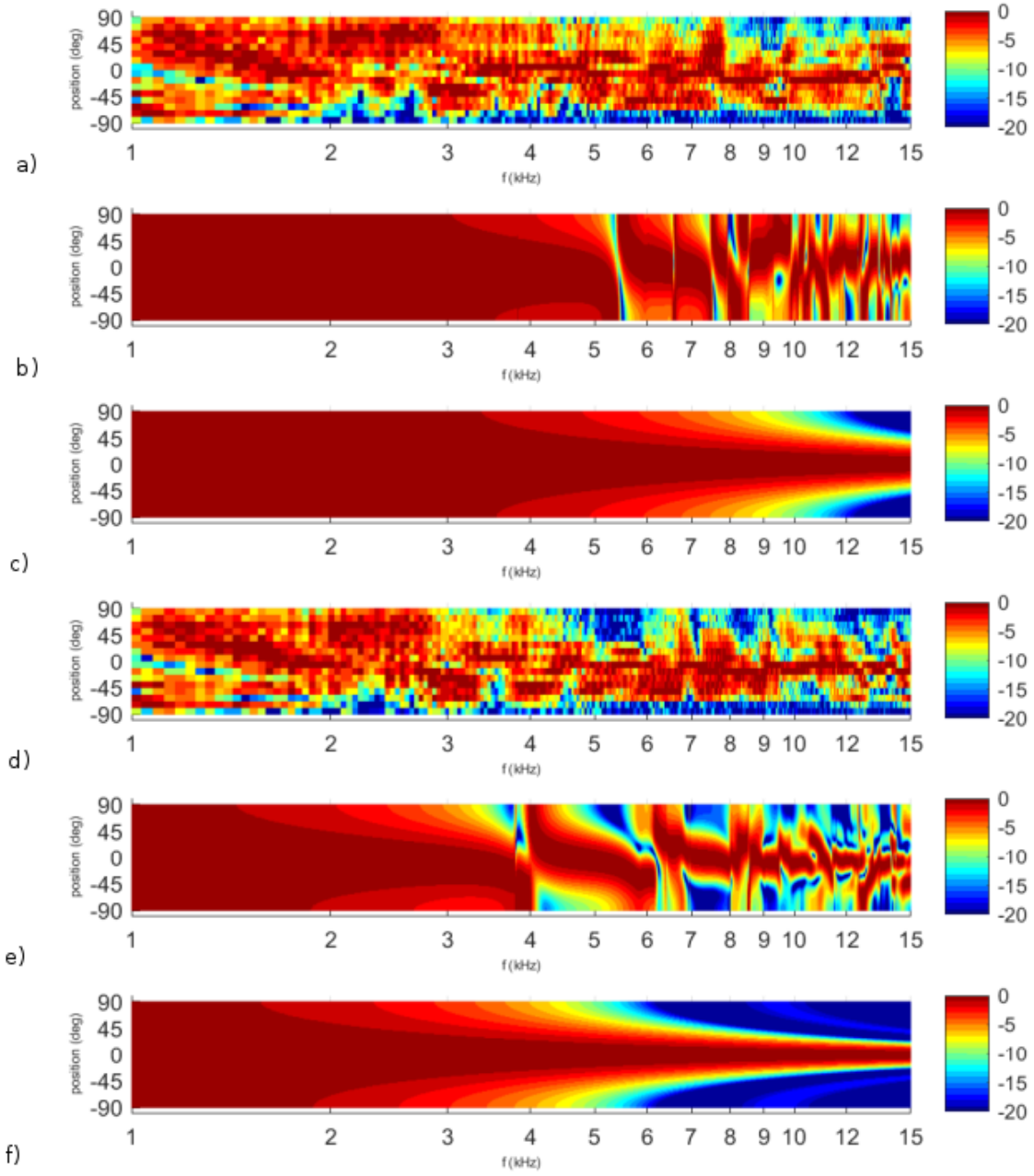


Figure 2. Normalized amplitude of the acoustic pressure radiated by a female professional singer and obtained with theoretical models as a function of the frequency and the angular position in the vertical plane (0°is located in front of the mouth). a) measurement with a normal mouth opening, b) simulation of a normal mouth opening based on the multimodal theory, c) simulation of a normal mouth opening using the vibrating piston model, d) measurement with a long mouth opening, e) simulation of a long mouth opening based on the multimodal theory and f) simulation of a long mouth opening using the vibrating piston model. No diffraction by the head and torso is taken into account in the simulations.

1798

The plane piston model explains the general trend of a beam narrowing as the frequency increases. It also describes the influence of the size of the mouth aperture on the main beam width: a large mouth aperture tends to generate more focused radiation patterns. This could affect the distance impression: the source seems to be nearer because less reverberation is heard. However, the plane piston model fails to explain the complexity of the patterns and the global downward direction of the beam above 2 kHz.

The multimodal method allows one to extend the plane piston model by adding the propagation of the higher order propagation modes inside the vocal tract and shows the influence on the radiation patterns. Taking into account higher order modes allows one to reproduce complex pattern variations, but only above the cut-off frequency of the first higher order modes (above 5 kHz and between 3-4 kHz for the normal and long configurations respectively). These complex patterns correspond to the radiation patterns of the higher order modes [17, 7]. The qualitative similarities between the measured patterns and the simulated ones (abrupt transitions and localized minima) show that the complexity observed experimentally can be, at least partially, attributed to higher order modes above 3-4 kHz. For the long configuration, regions of high and low levels attributable to higher order mode propagation occur around similar frequencies and angles. The patterns are not exactly similar, which was expected as the exact vocal tract geometry is not reproduced in the simulations.

An increase of mouth opening results in a shift of the effect of higher order modes to lower frequencies as well as an increase of their bandwidth. In fact, this is due to the larger vertical dimensions which reduce the cut-off frequencies of the higher order modes compared to the normal configuration.

The complexity of the measured radiation patterns below the cut-off frequency of the higher order modes can be explained by the diffraction by the head and the torso. Thus, the overall complexity of the measured radiation patterns probably results of a combination of the effect of the diffraction by the head and the torso and the radiation of higher order modes.

Since the experimental patterns of both mouth configurations are very similar up to 2 kHz, the effect of diffraction by the head and the torso is probably predominant in this interval. Above 2 kHz the influence of the mouth shape probably becomes more significant and starts to influence the radiation patterns:

- first through the width of the radiation beam, related to the size of the mouth opening. In fact, the deviation between both mouth configurations between 2-3 kHz is not predicted by the multimodal theory because the wavelength is still large compared to the mouth opening. Therefore, it is likely that the diffraction by the head of different width of radiation beam introduce unequal effects.
- Above the cut-off frequency of the first higher order modes, the mouth configuration substantially changes the radiation. Since higher order modes are influenced by the shape of the mouth opening, the internal shape of the mouth can potentially influence the radiation.

The fact that the a downward orientation of the beam can be reproduced by the multimodal simulation indicates that the shape of the mouth can potentially influence the main direction of radiation. However, the trend is not as clearly marked in the simulation as in the measurement. As the vocal tract of the singer has not been exactly replicated and merely approximated and no modeling of the lips nor the head and the torso are included, it was not expected to have an exact matching in the frequency interval of occurrence of the higher order modes. Still, the multimodal simulations predict that a larger mouth opening changes radiation drastically above 3.5 kHz.

A significant amount of energy is radiated in the 2 kHz to 4 kHz interval which also corresponds to the maximum of hearing sensitivity of the hear, and frequencies in the 8 kHz and 16 kHz octave bands can be relevant for the perception of singing quality [18]. On the other hand, such changes of beam width and direction of a sound source have been shown to be perceptually relevant [19, 20, 21]. Thus, the influence of the mouth configuration on directivity can potentially lead to perceptible effects. On the other hand, the higher order modes can increase the width of the beam and strongly modify its direction in some specific frequency intervals. This could be perceptible and play a role in the naturalness of speech and singing.

5 CONCLUSION

The measurement of the sound radiated by a professional singer shows that the use of two different mouth configuration leads to noticeably different radiation patterns. Thus, from 2 kHz on, a more focused radiation beam shifted downward is generated using a long mouth opening (which increases the vertical dimensions of the mouth).

The comparison of these measurements with simulations performed with higher order modes and a vibrating piston model shows that the general characteristics of the radiated sound can be estimated from the vibrating piston properties. Thus, the decrease of the width of the radiation beam at high frequency and with a greater mouth opening corresponds to the prediction of this model. However, it fails to predict the complex local variations of the directivity patterns as well as the global angular downward shift of the radiation beam in the case of the long mouth configuration.

On the contrary, complex local variations of the directivity patterns and downward shifts of the radiation beam are observed both experimentally and in the multimodal simulations. However, the global downward tendency is weaker. Accounting for the diffraction by the head and the torso and using a more realistic geometry could help to simulate in a more realistic way the radiation of speech and singing.

ACKNOWLEDGEMENTS

Part of this research was funded by the German Research Foundation (DFG), grant no. BI 1639/7-1.

REFERENCES

[1] Flanagan, JL. Analog measurements of sound radiation from the mouth, The Journal of the Acoustical Society of America, Vol 32, 1960, pp 1613–1620.

[2] Marshall, AH.; Meyer, J. The directivity and auditory impressions of singers, Acta Acustica united with Acustica, Vol 58 (3), 1985, pp 130-140.

[3] Chu, WT.; Warnock, ACC. Detailed directivity of sound fields around human talkers, Technical Report, Institute for Research in Construction (National Research Council of Canada, Ottawa ON, Canada), 2002, pp. 1–47.

[4] Katz, B.; d'Alessandro, C. Directivity measurements of the singing voice, International Congress on Acoustics (ICA 2007), 2007.,

[5] Monson, BB.; Hunter, EJ.; Story, BH. Horizontal directivity of low-and high-frequency energy in speech and singing, The Journal of the Acoustical Society of America, Vol 132 (1), 2012, pp 433-441.

[6] Kocon, P.; Monson, BB. Horizontal directivity patterns differ between vowels extracted from running speech, The Journal of the Acoustical Society of America, Vol 144 (1), 2018, pp EL7-EL12.

[7] Blandin, R,; Van Hirtum, A.; Pelorson, X.; Laboissière, R. The effect on vowel directivity patterns of higher order propagation modes, Journal of Sound and Vibration, Vol 432, 2018, pp 621-632.

[8] Kinsler, LE.; Frey, AR.; Coppens, AB.; Sanders, JV. Fundamentals of Acoustics, 4th edition, Wiley, 1999, p 189.

[9] Brandner, M.; Frank, M.; Rudrich, D. DirPat—Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers. Audio Engineering Society Convention 144, 2018.

[10] Nair, A.; Nair, G.; Reishofer, G. The low mandible maneuver and its resonential implications for elite singers, Journal of Voice, Vol 30 (1), 2016, pp 128.e13–128.e32.

1800

[11] Kob, M.; Jers, H. Directivity measurement of a singer, The Journal of the Acoustical Society of America, 1999, Vol 105 (2), pp 1003-1003.

[12] Brandner, M.; Sontacchi, A.; Frank, M.; Real-Time Calculation of Frequency-Dependent Directivity Indexes in Singing, DAGA 2019 Rostock.

[13] Aalto, D.; Aaltonen, O.; Happonen, RP.; Jääsaari, P,; Kivelä, A.; Kuortti, J.; Luukinen, JM.; Malinen, J.; Murtola, T.; Parkkola, R.; others Large scale data acquisition of simultaneous MRI and speech, Applied Acoustics, Vol 83, 2014, pp 64-75.

[14] Arnela, M.; Dabbaghchian, S.; Blandin, R.; Guasch, O.; Engwall, O.; Van Hirtum, A.; Pelorson, X. Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds, The Journal of the Acoustical Society of America, Vol 140 (3), 2016, pp 1707-1718.

[15] Xue, SA.; Hao, JG. Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry, Journal of Voice, Vol 20 (3), 2006, pp 391-400.

[16] Blandin, R.; Arnela, M.; Laboissière, R.; Pelorson, X.; Guasch, O.; Van Hirtum, A.; Laval, X. Effects of higher order propagation modes in vocal tract like geometries, The Journal of the Acoustical Society of America, Vol 137 (2), 2015, pp 832-843.

[17] Blandin, R.; Van Hirtum, A.; Pelorson, X.; Laboissière, R. Influence of higher order acoustical propagation modes on variable section waveguide directivity: Application to vowel [α], Acta Acustica united with Acustica, Vol 102 (5), 2016, pp 918-929.

[18] Monson, BB.; Lotto, AJ.; Story, BH. Detection of high-frequency energy level changes in speech and singing, The Journal of the Acoustical Society of America, Vol 135 (1), 2014, pp 400-406.

[19] Postma, BNJ.; Demontis, H.; Katz, BFG. Subjective evaluation of dynamic voice directivity for auralizations, Acta Acustica united with Acustica, Vol 103 (2), 2017, pp 181-184.

[20] Wendt, F.; Sharma, GK,; Frank, M.; Zotter, F.; Hoeldrich, R. Perception of spatial sound phenomena created by the icosahedral loudspeaker, Computer Music Journal, Vol 41 (1), 2017, pp 76-88.

[21] Frank, M.; Brandner, M. Perceptual Evaluation of Spatial Resolution in Directivity Patterns, DAGA 2019 Rostock.

2.5 A Pilot Study on the Influence of Mouth Configuration and Torso on Singing Voice Directivity

This work was published as:

M. Brandner, R. Blandin, M. Frank, and A. Sontacchi. A Pilot Study on the Influence of Mouth Configuration and Torso on Singing Voice Directivity. *Journal of the Acoustical Society of America*, 148(3):1169–1180, 2020. doi:10.1121/10.0001736.

The idea and concept of this article were outlined by me, the first author, with help from the second and the fourth author. I wrote the original draft of the manuscript with periodical contribu- tions from the second author. The revision and editing was done by me with help from the second, third, and fourth author. I did most of the programming, visualizations, measurements and prepared the data for publication. The second author provided the numerical analysis for the vocal tracts.

JASA ARTICLE

A pilot study on the influence of mouth configuration and torso on singing voice directivity[a)]

Manuel Brandner,[1,b)] Remi Blandin,[2] Matthias Frank,[1] and Alois Sontacchi[1]
[1]Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz 8010, Austria
[2]Institute of Acoustics and Speech Communication, Technical University of Dresden, Dresden 01062, Germany

ABSTRACT:
Directivity of speech and singing is determined primarily by the morphology of a person, i.e., head size, torso dimensions, posture, and vocal tract. Previous works have suggested from measurements that voice directivity in singing is controlled unintentionally by spectral emphasis in the range of 2–4 kHz. The attempt is made to try to identify to what extent voice directivity is affected by the mouth configuration and the torso. Therefore, simulations, together with measurements that investigate voice directivity in more detail, are presented. Simulations are presented for a piston in an infinite baffle, a radiating spherical cap, and an extended spherical cap model, taking into account transverse propagation modes. Measurements of a classical singer, an amateur singer, and a head and torso simulator are undertaken simultaneously in the horizontal and vertical planes. In order to assess differences of voice directivity common metrics, e.g., horizontal and vertical directivity indexes, are discussed and compared to improved alternatives. The measurements and simulations reveal that voice directivity in singing is affected if the mouth opening is changed significantly. The measurements show that the torso generates side lobes due to diffraction and reflections at frequencies related to the torso's dimensions.
© 2020 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1121/10.0001736
(Received 18 December 2019; revised 25 June 2020; accepted 24 July 2020; published online 4 September 2020)
[Editor: Vasileios Chatziioannou] Pages: 1169–1180

I. INTRODUCTION

Voice directivity describes how sound is radiated into space from the mouth and/or nasal orifices of a person. Voice radiation is characterized by its directionality and the efficiency with which the sound produced by the vocal tract is transmitted outside. Voice directivity patterns can be calculated from simulated or measured sound pressure levels at fixed distances from a defined center position along a circle or sphere around a subject. The most prominent influences on voice directivity (Blandin *et al.*, 2015; Blandin and Brandner, 2019; Blandin *et al.*, 2018; Cabrera *et al.*, 2011; Chu and Warnock, 2002; Flanagan, 1960; Katz and D'Alessandro, 2007; Marshall and Meyer, 1985) can be summarized as follows:

- shape and size of head and torso (fixed),
- posture, head inclination (variable),
- vocal tract geometry and mouth opening (variable), and
- spectral emphasis due to vocal technique (variable).

A recent review of research on voice directivity can be found in Abe (2019). While there are several studies in the literature that investigate the directivity patterns of human speech, only a few reveal to which extent a person can control or at least incidentally change these patterns. In the studies of Cabrera (2004), Cabrera *et al.* (2011), and Chu and Warnock (2002), general directivity properties of speakers and singers have been investigated. The data are calculated over longer time frames of running speech or sung phrases. The pioneering work of Marshall and Meyer (1985) is also noteworthy. They analyze three vowels for three singers but present the results solely in octave bands and not in full detail. Changes of directivity in detail (third-octave bands or higher resolution) for different phonemes of speakers have been recently presented in Kocon and Monson (2018) and, to the author's knowledge, for a classical singer only in Katz and D'Alessandro (2007). In the latter, the singer's directivity has been investigated in detail by calculating long-term averaged spectra (LTAS) of specific vowels rather than sung phrases, but the corresponding mouth openings have not been discussed. The directivity patterns show subtle differences between different vowels and performance styles for one test subject. Such detailed results of steady conditions cannot easily be obtained by LTAS from running speech because coarticulation can affect the mouth opening for a specific phoneme and, therefore, smear the results.

Unconsidered in all these studies is that for the geometry of the vocal tract and simpler cylindric geometries, higher order modes (transverse propagation modes) can occur. Transverse propagation modes generate a nonuniform velocity field on the mouth exit plane, which departs from the plane piston assumption (Savkar, 1975; Snakowska *et al.*, 1996). Recently published data in Blandin *et al.*

[a)]This paper is part of a special issue on Modeling of Musical Instruments.
[b)]Electronic mail: brandner@iem.at

https://doi.org/10.1121/10.0001736 JASA

(2018) reveal that the geometry of the vocal tract can have influence on the sound radiation above 6 kHz in speech production.

In general, an increase in directivity is expected if larger aperture diameters are used, which correspond to larger mouth openings. A significant increase or decrease of voice directivity would introduce a change of the perceived source distance at the listeners position (Wendt *et al.*, 2017).

The influence of a professional singing technique on the vocal tract configuration has been shown by the use of ultrasound equipment in Nair *et al.* (2016). In this study, it is assumed that as a strategy a larger mouth opening is used by professional singers. If changing the oral configuration in singing is related to optimizing the vocal output, then measurement results of differences in directivity for different mouth openings are supposed to relate to this optimization process.

An initial objective of this work is to identify if simple models can serve as good approximations of human voice directivity. A second objective is to determine the effect of the torso on voice directivity and in which frequency region the effects are most prominent. However, the main objective of the work is to determine if different mouth openings while vocalizing the German vowel /a/ show an effect on voice radiation in singing. The paper is structured as follows. First, we present the three used simulation models, our measurement setup, data visualizations, and metrics in Sec. II. Second, we compare the results for the three used sound radiation models in Sec. III A. Third, we present the measurement results to investigate the effect of the torso and different mouth openings in Sec. III B. Finally, on the basis of our results, we discuss how the head, torso, and mouth configuration influences voice directivity in Sec. IV. For an objective interpretation and comparison of the results, we use common metrics and introduce a new measure, the so-called *energy vector* as an enhanced descriptor.

II. METHODS

In this section, we present a wide variety of opportunities to achieve an objective analysis and reproducible measurement results for singing voice radiation and directivity analysis.

A. Simulations

We investigate currently known simulation models on their applicability to provide insights on influences from diffraction or higher order modes on voice directivity. This should allow us to better understand specific effects seen in voice directivity measurement data. Therefore, we investigate the piston model, which accounts solely for different mouth opening areas with a uniform particle velocity distribution, the spherical cap model, which includes also head diffraction and also a uniform particle velocity distribution, and the extended spherical cap model, which also accounts for the propagation of transverse modes, i.e., the superposition of uniform and nonuniform particle velocity distributions.

1. Piston model

As described in Flanagan (1960), the mouth opening can be approximated in the simplest case by a spherical source on a spherical baffle, where the latter corresponds to the human head. Furthermore, he proposes a small piston instead of a spherical source. The far-field pressure for a circular baffled piston with a radius r_p at a radial distance r is given as follows:

$$p(kr,\vartheta) = \frac{2J_1(kr_p\sin\vartheta)}{kr_p\sin\vartheta}\frac{i\rho_0 ck}{2r}r_p^2\tilde{v}e^{i(\omega t-kr)}, \tag{1}$$

with J_1 denoting the first-order Bessel function, the angular wave number k, the density of air ρ_0, the speed of sound c, and the membrane velocity $\tilde{v}$. The sound pressure is dependent on the distance r and due to axial symmetry only on the angle ϑ. Therefore, a simple calculation regarding the influence of mouth openings corresponding to different radii of a piston is feasible. Directionality increases noticeably for $k(d/2) \gg 1$, whereas the frequency of the transition from a nondirectional to directional radiation pattern for a piston—which will be called the transition frequency within the rest of the paper—can be calculated with $f \approx c/(\pi d)$ (Zwicker and Zollner, 1993).

2. Spherical cap model

An even more realistic approach to simulate the sound radiation from a person's mouth is given by the spherical cap model described in Zotter and Frank (2019). In contrast to the piston model, the sound is radiated by a spherical cap located on a rigid sphere with a radius r_0 and an aperture angle α. The model does allow one to simulate sound radiation for angles larger than $\pm 90°$. The pressure distribution in the far-field for a single spherical cap can be calculated by the use of the Legendre polyomials P_n of order n as follows:

$$p(kr,\vartheta) = \frac{\rho_0 c}{i4\pi}\sum_{n=0}^{\infty}(2n+1)\frac{h_n^{(2)}(kr)w_n}{h_n'^{(2)}(kr_0)}P_n(\cos\vartheta), \tag{2}$$

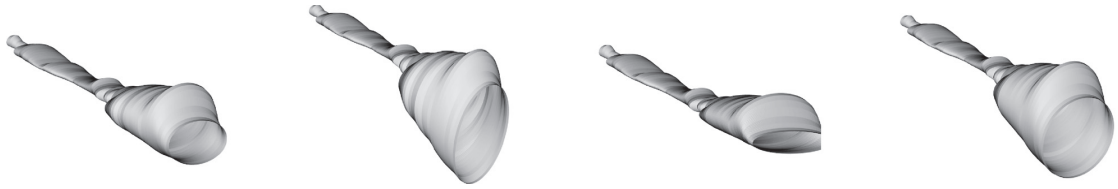
where r is the radius in space, $h_n^{(2)}(x)$ is the spherical Hankel function of the second kind, and $h_n'^{(2)}(x)$ is its first derivative. The aperture opening is accounted for by the use of a weighting function w_n (Zotter and Frank, 2019),

$$w_n = \begin{cases} \cos\left(\frac{\alpha}{2}\right)P_n\left[\cos\left(\frac{\alpha}{2}\right)\right] - P_{n-1}\left[\cos\left(\frac{\alpha}{2}\right)\right], & n>0, \\ 1-\cos\left(\frac{\alpha}{2}\right), & n=0. \end{cases} \tag{3}$$

3. Extended spherical cap model

The plane wave assumption inside the vocal tract does not hold for frequencies higher than about 4 kHz because

JASA
https://doi.org/10.1121/10.0001736



(a) Normal (b) Long (c) Wide (d) Open

FIG. 1. Vocal tract geometries used for the transverse propagation mode simulations. The upper left extremity of each geometry corresponds to the vocal fold location and the lower right extremity of each geometry corresponds to the mouth opening. The mouth opening is facing towards the 0° direction.

given the dimensions of the vocal tract, the transverse propagation modes can potentially propagate from this frequency on (Blandin *et al.*, 2015). Thus, the particle velocity on the mouth exit may become nonuniform, and the plane piston and the spherical cap models are no more valid. Here, the spherical cap model is extended by accounting for the variations of the particle velocity. The particle velocity field over the mouth exit is discretized in N_p point sources of acoustic flow velocity amplitudes ν_l. The acoustic pressure is obtained as the summation of the contributions of each point source,

$$p(kr,\theta) = \frac{\rho_0 c}{i4\pi N_p}\sum_{l=1}^{N_p}\sum_{n=0}^{\infty}(2n+1)\frac{h_n^{(2)}(kr)}{h_n'^{(2)}(kr_0)} \times P_n(\cos\,\gamma_l)\nu_l\Delta S_l, \quad (4)$$

where γ_l is the angular distance between the coordinates of the source point θ_l and the coordinates of the reception point θ and ΔS_l are the equivalent surface areas. The particle velocity field at the mouth exit is calculated by using the *multimodal* method, which relies on the projection of the acoustic field on the propagation modes of a locally uniform waveguide. As this approach includes the plane mode, it can be considered as an extension of the plane wave theory with higher order modes. The *multimodal* theory used in this work is described in Blandin *et al.* (2015) and Blandin *et al.* (2016). Similar results may be obtained by the use of any other three-dimensional (3D) acoustic simulation method like the finite element or finite difference method.

In Fig. 1, the vocal tract geometries used for the simulations are shown. The vocal tract configurations are defined as "normal," "long," "wide," and "open" for the ease of reading and correspond to the mouth openings shown in Fig. 2. However, we recognize that the mouth openings are not independent of the vocal tract configurations. Furthermore, the mouth opening and its corresponding vocal tract configuration will affect the formant characteristics, i.e., their frequency, bandwidth, and strength. This will receive no further attention here as our focus lies on the sound radiation patterns. The used approximations of the real vocal tract of the singer are adaptations of a 3D geometry extracted from a magnetic resonance image for the vowel /a/ of a male speaker (Aalto *et al.*, 2014; Arnela *et al.*, 2016). The length and the volume of the oral and pharyngeal cavity of the original geometry were adapted to female dimensions by utilizing the data of 120 male and female subjects (Xue and Hao, 2006). An elliptical mouth opening is used in which the height and width of the mouth openings have been taken from the female classical singer participating in this study (cf. Table II). For a more detailed explanation of the geometry design, see Blandin and Brandner (2019).

B. Measurements

1. Test subjects and conditions

One professional classical singer with a master's degree in classical voice and international singing experience and an amateur singer are investigated in the study. The classical singer and amateur singer were both instructed to orient their head always in the forward direction. The mean head inclination angle acquired by utilizing a tracking system validates that only slight movements did take place; see Table I.

As the main goal of the study is the analysis of the influence of the mouth opening on voice directivity, four mouth openings are investigated (Table II; Fig. 2).

As a target vowel, we define the German vowel /a/ even though slight deviations are expected to occur due to a change of mouth opening. The classical singer is asked to maintain the same vocal effort for all conditions and merely change the mouth opening.

To investigate the influence of the upper body on the radiated sound pressure field, measurements are made with a Brüel and Kjaer 4128 head and torso simulator (HATS; Nærum, Denmark). As an excitation signal, we use a logarithmic

(a) Normal (b) Long (c) Wide (d) Open

FIG. 2. Images of the mouth openings (a) *normal* (like in speech), (b) *long* (lowered jaw), (c) *wide*, and (d) *open* (lowered jaw and large horizontal mouth width) of the classical singer.

https://doi.org/10.1121/10.0001736 JASA

TABLE I. Mean head inclination angle $\bar{\vartheta}$ acquired from tracking data for all four conditions for the classical singer.

	Normal	Long	Wide	Open
$\bar{\vartheta}$ (deg)	1.7	5.8	1.3	6.6

sweep. The torso influence in the horizontal and vertical directivities can be discussed for the HATS as the head can be removed from the torso and measured separately. The mouth opening has a horizontal width of 3 cm and a vertical distance of 1 cm.

2. Room conditions

Measurements were carried out in a sound treated measurement room with absorptive material on the walls and floor at the Institute of Electronic Music and Acoustics. The mean room reverberation between 400 Hz and 1 kHz is below 75 ms and above 1 kHz below 50 ms. The volume of the room is approximately 50 m^3 with a floor area of 22.50 m^2.

3. Double circle microphone array (DCMA)

The measurements of source radiation patterns are undertaken using a microphone array with a radius of 1 m consisting of two circular rings, one placed in the horizontal plane and the other one placed in the vertical plane, described in Brandner *et al.* (2018). Each of the rings can hold up to 32 microphones (NTI m2230, Schaan, Liechtenstein)—which means an angular spacing of 11.25°—resulting in a maximum number of 62 microphones as both rings intersect in the front and back of the array. In addition, a reference microphone is used and is located at the exact center of the microphone array.

4. Measurement procedure

In order to facilitate reproducible and comparable data, the head of the performer is equipped with reflective markers for optical tracking. The singer is asked to sit within the measurement setup as close as possible to a centered reference microphone. A visual feedback of the position is provided, whereby the mouth is defined as the acoustical center. The performer is asked to sing a glissando, starting at a low pitch (G4, 392 Hz) and ending one octave above the starting pitch. The glissando ensures that a wide range of frequencies are captured; see Kob and Jers (1999). Impulse responses for vocalized phonemes from directivity measurements with a reference signal can be acquired, which leads to a similar approach as the exponential swept-sine method in Farina (2000) to further improve the signal-to-noise ratio. The impulse responses are cut to a length of 1024 samples at a sampling frequency of 44.1 kHz, which results in a frequency resolution of 43 Hz. Simultaneous array measurements as outlined in Sec. II B 3 reduce both measurement time and positioning drifts.

TABLE II. Dimensions of the mouth openings of the classical singers for the four investigated mouth configurations. Measures are estimated from pictures made during the measurements (Fig. 2). The dimensions are used for the mouth exit of the vocal tract geometries used in the simulations (Fig. 1).

	Normal	Long	Wide	Open
Width (cm)	4.8	4.2	6.6	5.9
height (cm)	2.7	6.6	1.4	5.2

5. Measurement uncertainty

The variability of the measurement results due to positioning errors has been investigated with the HATS. The HATS is positioned with various radial offsets from the center position (up to 11 cm off-center) of the measurement setup. Therefore, we expect for the largest positioning error of 11 cm radially a maximum magnitude error of ±1 dB. The standard deviations of the horizontal directivity index (HDI) and the vertical directivity index (VDI) for all tested positioning offsets lie below 1 dB in the vicinity of 1.5 kHz and below 0.6 dB elsewhere. The positioning error shows minimal influence on the calculated directivity indexes and can be considered neglectable if the positioning error stays at least below a magnitude of 5 cm.

C. Metrics

In addition to the visualization of voice directivity in polar plots and normalized acoustic pressure maps, the following metrics will be used to investigate differences. The metrics are calculated from third-octave smoothed data.

1. Directivity index for a single plane

In accordance to literature (Cabrera, 2004; Tylka *et al.*, 2015), we define the directivity factor $\gamma_p(\omega) = P_{\text{on-axis}}/P_{\text{mean}}$ for the horizontal and vertical planes. It is defined by the ratio of the on-axis power $P_{\text{on-axis}}$ to the average power P_{mean} of all sampling positions on the respective plane. The HDI and VDI evaluated at an angular frequency ω for each plane are then defined in dB as follows:

$$\mathrm{DI}(\omega) = 10\ \log_{10}(\gamma_p(\omega)). \tag{5}$$

2. Front-to-back ratio

The HDI and VDI tend to decrease if side lobes with higher levels than the on-axis level occur, which can lead to the false conclusion that at these frequencies, the radiation pattern gets more omnidirectional. We define the front-to-back ratio (FBR) in dB as the ratio of the average power radiated to the front P_{front} and to the back P_{back} [Eq. (6)].

$$\mathrm{FBR}(\omega) = 10\ \log_{10}\frac{P_{\text{front}}(\omega)}{P_{\text{back}}(\omega)}. \tag{6}$$

3. Upward-to-downward ratio

This metric is used to investigate whether most of the energy in the vertical plane is radiated upward or rather

JASA
https://doi.org/10.1121/10.0001736

downward to the floor. The upward-to-downward ratio (UDR) denotes the ratio of how much power is radiated to the upper half space P_{upper} versus the energy radiated to the lower half space P_{lower} in dB:

$$\text{UDR}(\omega) = 10\,\log_{10}\frac{P_{\text{upper}}(\omega)}{P_{\text{lower}}(\omega)}. \tag{7}$$

4. Energy vector

The energy vector $\boldsymbol{r_E}$ in Eq. (8) can be utilized to describe the direction and the width of the main lobe of an acoustic source (Fig. 3). This measure, first introduced by Gerzon (1992), is commonly used in the context of 3D loudspeaker setups and their evaluation (Frank, 2013) but is useful in the description of properties of any arbitrary sound source radiation:

$$\boldsymbol{r_E} = \frac{\sum_{i=1}^{L} |H(\omega,\phi_i)|^2 \boldsymbol{m_i}}{\sum_{i=1}^{L} |H(\omega,\phi_i)|^2}. \tag{8}$$

The frequency dependent magnitudes $H(\omega,\phi_i)$ are multiplied by the vectors $\boldsymbol{m_i} = [\cos(\phi_i),\ \sin(\phi_i)]^T$ of each measurement position i, $i = 1, 2, \ldots, L$ in each respective plane and normalized by the sum of the energy, yielding a normalization of the vector between the limits 0 (omnidirectional) to 1 (maximum focus to one direction). The following two metrics will be used: (i) the main beam width in each plane $\theta_w = 2\arccos||\boldsymbol{r_E}||$ and (ii) the main direction in the vertical plane θ_s.

D. Visualization

1. Polar patterns

A quasi-continuous representation of arbitrary radiation directions for the azimuth angle φ and elevation angle ϑ can be rendered from given discrete measurement positions by applying the circular or spherical harmonics (SH) transform (Zotter, 2009). The polar patterns are displayed logarithmically and show a dynamic of 25 dB in each plane. The data presented in Figs. 9 and 11 are third-octave smoothed. The visualization and analysis tools are freely available within the DirPat-project, which is in the spirit of open data and reproducible research and discussed in Brandner *et al.* (2018).

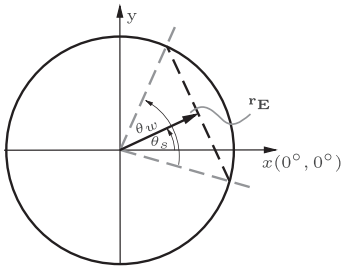


FIG. 3. Schematic of the energy vector and its corresponding source angle θ_s and source width θ_w.

2. Acoustic pressure maps

The acoustic pressure maps show the amplitude of the acoustic pressure normalized by the maximum over the angular position for each frequency and they are limited to a dynamic range of 20 dB. We use a frequency resolution of 10 Hz for the displayed data of the simulations, whereas for the measurement data, a frequency resolution of 43 Hz is used, and for visualization, semitone-octave smoothing is used.

III. RESULTS

A. Simulation results

1. Piston model

Normalized acoustic pressure maps are shown in Figs. 4(a) and 4(b) for a piston diameter of $d = 3$ cm and

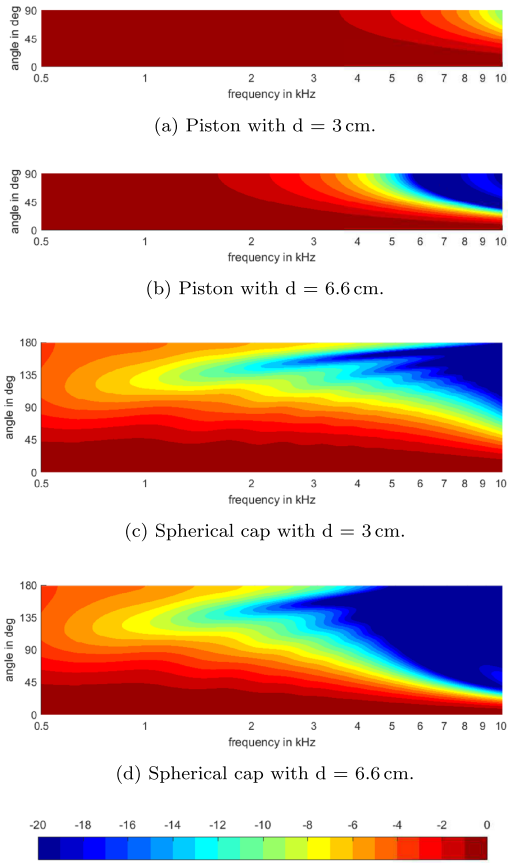


FIG. 4. (Color online) Normalized acoustic pressure as a function of the angular position and frequency for the piston [(a), (b)], and the spherical cap model [(c), (d)] for aperture diameters of 3 and 6.6 cm. In contrast to the piston model, sound radiation for the spherical cap model can be calculated for the front and the back.

https://doi.org/10.1121/10.0001736

JASA

$d = 6.6$ cm, respectively, corresponding to the width of the mouth opening of the HATS and the classical singer using the *wide* mouth opening configuration. An increase of directionality for $d = 3$ cm and $d = 6.6$ cm is visible at around 3.6 kHz and 1.7 kHz, respectively, and increases further with frequency, which is indicated by a steady decrease of the amplitudes toward the side. Due to the axial symmetry of the piston, only half of the acoustic pressure maps are shown. As the piston is located in an infinite baffle, the sound pressure levels are only calculated in the half space (i.e., only in front of the baffle).

2. Spherical cap model

The size of the sphere is set to the average size of a human head ($r_0 = 8.5$ cm). The model allows one to calculate the directivity in the full-space. The spherical cap model shows a narrower directivity pattern toward the front, already at lower frequencies, in contrast to the piston model in the infinite baffle [see Figs. 4(c), and 4(d)]. The amplitudes at 90° for 3.6 and 1.7 kHz are of around 3.5 dB less than for the piston model. The plots also show larger amplitudes at 180°. This is due to constructive interference for frequencies with a wavelength of the same order of size as the sphere radius.

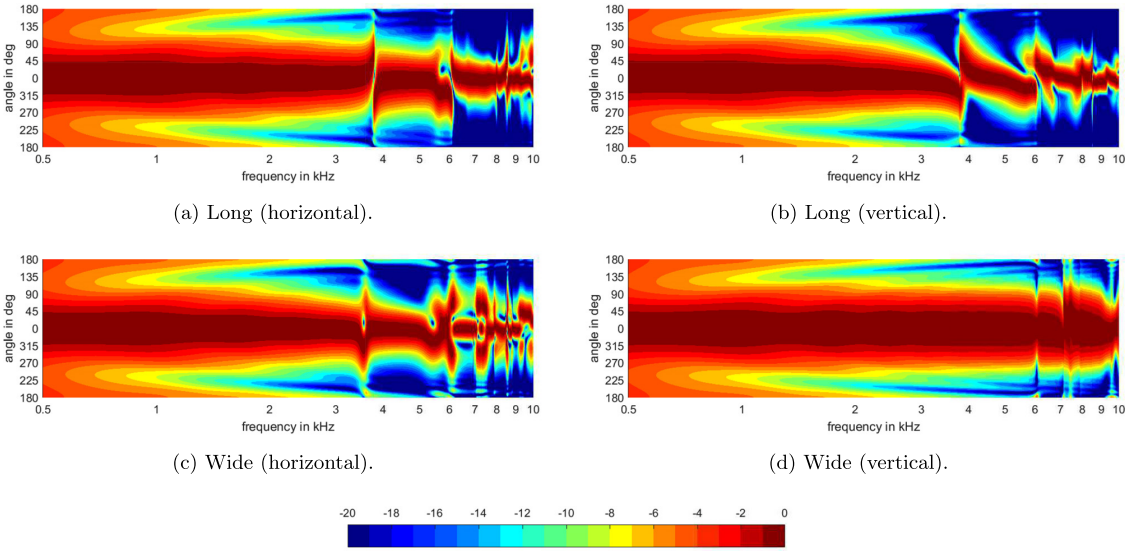
3. Extended spherical cap model

The directivity maps for an extended spherical cap model by accounting for transverse propagation modes are presented for the *long* and *wide* mouth opening configurations in the horizontal and vertical planes in Fig. 5. The effect of the transverse propagation modes is visible as vertical streaks, which can already propagate from 3.6 kHz on; see Fig. 5(c). The cut-on frequency of the first transverse propagation mode increases to 3.8 kHz as the mouth width is decreased [Fig. 5(a)]. In the horizontal plane, the *long* configuration shows a broader main beam width between 2.5 and 6 kHz with almost the same amplitude between 270° and 90° [Fig. 5(a)] in comparison to the *wide* configuration in Fig. 5(c), in which the main beam width is narrower. Furthermore, the transverse propagation modes introduce asymmetries around their cut-on frequencies, which can be seen around 3.6 and 6 kHz. Above 6 kHz, less transverse propagation modes occur for the *long* configuration. In contrast, the vertical plots in Figs. 5(b) and 5(d) show much larger differences. In general, we see that the *wide* configuration has a much broader main beam width. In Fig. 5(b), we see that at and above 3.8 kHz, the first transverse propagation mode affects the directivity of the *long* configuration by creating a local maximum around 90°, whereas for the *wide* configuration the first transverse propagation mode occurs at 6 kHz with almost no impact; see Fig. 5(d). The transverse propagation modes affect the main beam direction and the sound is, in general, directed downward above 3.8 kHz for the *long* configuration. For all configurations, sound radiation gets more complex in comparison to the piston and spherical cap model, showing significant variations in frequency intervals on the order of 100 Hz. As the frequency and size of the mouth opening increase, the number of transverse propagation modes increases. Further data on the *normal* and *open* mouth configurations can be found in the supplementary material.[1]

B. Measurement results

1. Torso influence (HATS)

The normalized acoustic pressure maps for the HATS with and without torso are shown in Figs. 6(a) and 6(b),

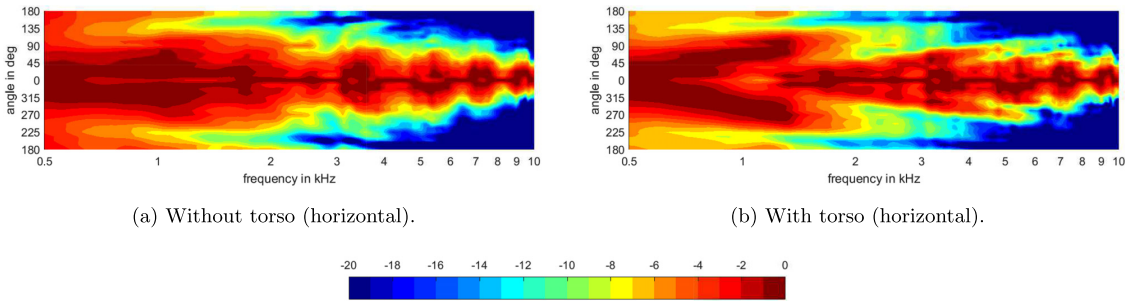


(a) Long (horizontal). (b) Long (vertical).

(c) Wide (horizontal). (d) Wide (vertical).

FIG. 5. (Color online) Normalized acoustic pressure as a function of the angular position and frequency for the simulations accounting for transverse propagation modes. (a), (b) The *long* mouth configuration and [(c), (d)] the *wide* mouth configuration are shown in the horizontal plane and vertical plane, respectively.

JASA
https://doi.org/10.1121/10.0001736



(a) Without torso (horizontal). (b) With torso (horizontal).

FIG. 6. (Color online) Normalized acoustic pressure for the HATS as a function of the angular position and frequency (semitone-octave smoothed). The HATS is measured (a) with and (b) without torso.

respectively. The normalized acoustic pressure map of the HATS without torso shows that the amplitudes below 2 kHz decrease by 1.5 dB already for angles larger than 45° (or lower than 315°) except at around 1.1 kHz. Furthermore, sound is also radiated toward the back for frequencies below 2 kHz with lower amplitudes between 90° and 135° and higher amplitudes at around 180° for frequencies below 1 kHz. The main beam width toward the front decreases as the frequency increases; see Fig. 6(a). However, it broadens in specific frequency bands around 3.3, 5.3, 7.3, and 9.3 kHz. It can be seen that sound radiation looks very similar—except for the broadening of the main beam width at the mentioned frequencies—if compared to the results of the spherical cap model for an aperture size of 3 cm.

We turn now to the measurement results for the HATS including the torso. Smaller amplitudes are visible toward the sides and the back already at lower frequencies in the range of 0.5–1 kHz. The most striking result to emerge from the data is that the torso provokes distinct side lobes within the range of 1–2 kHz, 3–5 kHz, and less pronounced within 6–7 kHz. A decrease of amplitudes toward the front on the order of −5 dB is visible at around 1.1 kHz. We now use the HDI and VDI metrics to compare the measurements. The HDI and VDI only quantify the directionality in the corresponding plane. Figure 7 shows that the strongest deviations in the range of 3 dB and 2 dB occur around 1.1 kHz and 2 kHz, respectively. The differences are smaller than 1 dB above 2 kHz for the horizontal plane, and the differences are smaller than 1 dB above 2.5 kHz for the vertical plane. Furthermore, the ripples visible in the acoustic pressure maps are also visible as drops of the HDI and VDI values at the same frequencies. The directivity index decreases to negative values around 1 kHz due to strong side lobes provoked by the torso.

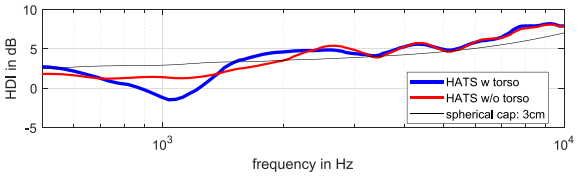
C. Comparison of test subjects and HATS

In this section, we give a qualitative comparison of the directivity of the HATS with torso, the classical singer, and the amateur singer. For the comparison, we use the measurements of the test subjects with the *normal* mouth opening configuration.

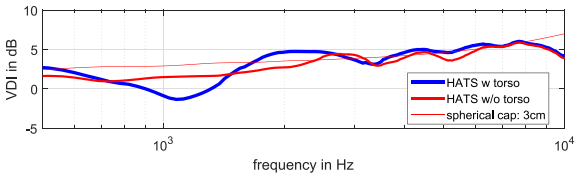
In Fig. 8(a), the HDI values show a general tendency to increase over frequency for the test subjects and the HATS. Although, the curves also show drops at 1.1, 3.3, and 5.1 kHz for the HATS and at 750 Hz, 1.7 kHz, 3 kHz, and 8 kHz for the amateur singer, and at 1, 2.2, and 5.2 kHz for the classical singer. These drops also occur for the VDI values at similar frequencies in Fig. 8(b). The VDI curves show a slight increase over frequency for the HATS and the classical singer and a decrease for the amateur singer above 4 kHz.

Polar patterns for the test subjects and the HATS show greatest accordance if they are evaluated at the frequencies corresponding to the identical minima of the directivity indexes (DIs) than at the same frequencies. In order to illustrate the dependency of the side lobes on the torso dimensions, the directivity patterns of the HATS, the amateur singer, and the classical singer are plotted at frequencies corresponding to the same first minimum of the DI curves in Fig. 9, that is, 1.1 kHz, 800 Hz, and 950 Hz, respectively (Fig. 9).

The FBR and UDR values are presented in Figs. 8(c) and 8(d), respectively. The FBR generally increases with



(a) HDI.

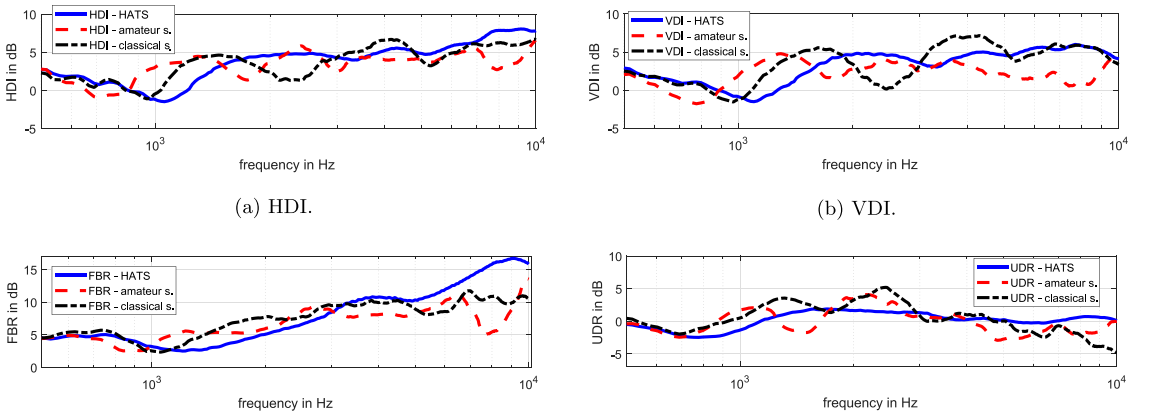


(b) VDI.

FIG. 7. (Color online) (a) HDI and (b) VDI for the HATS with torso (thick), without torso (medium), and the spherical cap for $d = 3$ cm (thin). The HDIs and VDIs are calculated from third-octave smoothed frequency data.

https://doi.org/10.1121/10.0001736

JASA



(a) HDI. (b) VDI.

(c) FBR. (d) UDR.

FIG. 8. (Color online) (a) The HDI and (b) VDI values, as well as (c) FBR and (d) UDR values are shown for the HATS (solid), the amateur singer (dashed), and the classical singer (dashed-dotted).

frequency for the test subjects and the HATS. Prominent drops occur within the region of 800 Hz–1 kHz. The directionality of the HATS increases for frequencies above 7 kHz. The FBR values above 7 kHz stay at 10 dB for the classical singer and decrease for the amateur singer. Largest deviations for the UDR values of about 5 dB occur at 1.5 kHz between both the test subjects and the HATS around 1.5, 2.5, and above 8 kHz.

D. Influence of the mouth configuration on radiation

The HDI, VDI, FBR, and UDR values obtained with the four mouth configurations on the classical singer are presented in Figs. 10(a) and 10(b). The deviation of HDI between the different configurations is up to 2.6 dB in a broad frequency range of 1.5–3 kHz and within the narrower range of 7–8 kHz. The VDI values also differ in the frequency range of 1.5–3 kHz and within the range of 4–7 kHz. The FBR values in Fig. 10(b) show more prominent differences with deviations of up to 5 dB within the range of 1.5–3 kHz, whereas the *wide* configuration shows the highest FBR. The UDR values show that along the vertical axis up to 1.8 kHz, the sound is focused toward the same direction for all mouth openings. Differences of the magnitude of 3 dB occur within the region of 2–3 kHz and above 4 kHz. The UDR values decrease to −5 dB for the *long* configuration, indicating that most of the energy is radiated downward from 4 kHz on. A similar decrease is shown for the *open* configuration but in the smaller range of 4–6 kHz.

In Fig. 11, the polar patterns show the directivity in more detail at 2.4 and 5 kHz. We see larger differences between mouth openings in the magnitude of 5 dB at 2.4 kHz in the horizontal plane and in the magnitude of almost 10 dB at 5 kHz in the vertical plane. At 5 kHz, the *normal* and *wide* mouth openings distribute the energy almost symmetrically in both planes. This agrees with the features highlighted by the FBR and UDR metrics.

Let us now turn to how we can gain the abovementioned information by using the angular information of the energy vector in our detailed analysis. In Figs. 12(a)–12(d), we present the beam width θ_w for each plane and the main direction of sound θ_s for the vertical plane in degrees for each proposed mouth opening. The *wide* and *open* conditions show the highest directionality in the horizontal plane within 1.5–4 kHz. In the vertical plane, the *long* and *open* conditions from 5 to 9 kHz are most directional. In addition, the *open* condition focuses sound between 1.5 and 4 kHz and between 5 and 9 kHz in both planes and tends, therefore, to be the most directional condition. The larger mouth openings *long* and *open* tend to radiate most of the energy toward the floor within 3–5 kHz. This focusing of sound toward the floor occurs also at higher frequencies for the *long* condition, but this tendency decreases as the frequency increases further. Finally, we present a comparison of the acoustic pressure maps for the classical singer for the *long* and *wide* conditions in Fig. 13 in the horizontal and vertical planes.

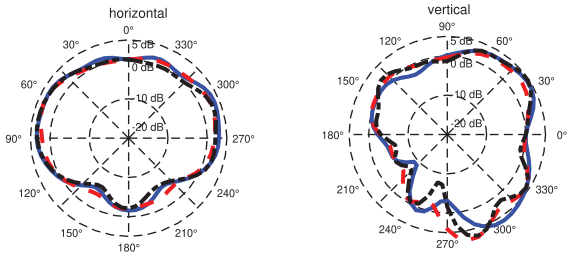
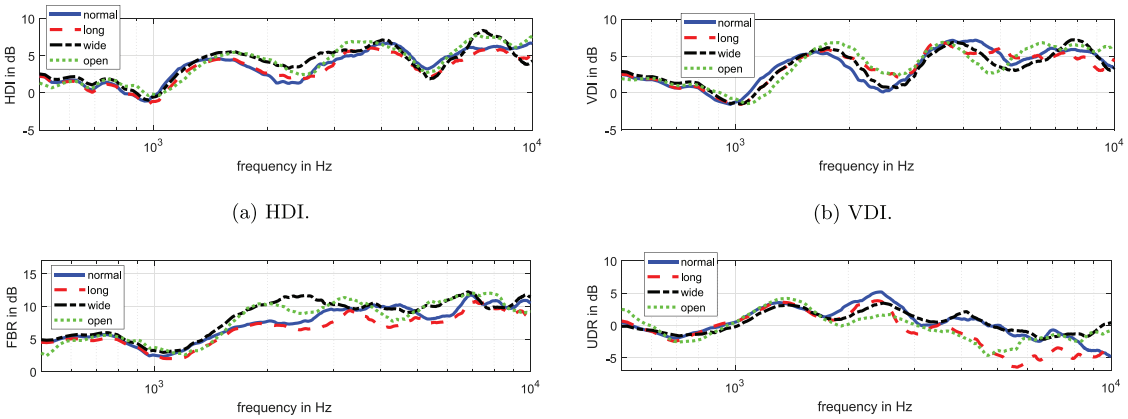


FIG. 9. (Color online) Directivity patterns measured on the HATS (solid blue), the amateur singer (dashed red), and the classical singer (dashed-dotted black) for *normal* mouth configuration are compared at the frequency of the corresponding first valley of the respective HDI/VDI curve shown in Fig. 8.

JASA
https://doi.org/10.1121/10.0001736



(a) HDI. (b) VDI.

(c) FBR. (d) UDR.

FIG. 10. (Color online) The HDI and VDI, the FBR, and the UDR measured on the classical singer with the four mouth openings investigated: *normal* (solid), *long* (dashed), *wide* (dashed-dotted), and *open* (dotted); see Fig. 2.

The acoustic pressure maps show similar side lobes as in the results for the HATS with torso [cf. Fig. 6(b)]. Although in the higher frequency region, more complex patterns are visible. Due to a change of mouth opening from *long* to *wide*, the most affected frequency region is above 1.5 kHz. However, all conditions show quite complex sound radiation above 2 kHz. Several local minima can be seen for the *long* and *wide* configurations in both planes. In particular, we see local minima in both planes for the long configuration at 3.5 kHz and for the wide configuration at 4.5 and 8 kHz. In the horizontal plane, the patterns of the wide configuration are more complex above 4 kHz than the patterns of the long configuration. In general, the *wide* mouth configuration tends to show a narrower pattern from 1.5 kHz on.

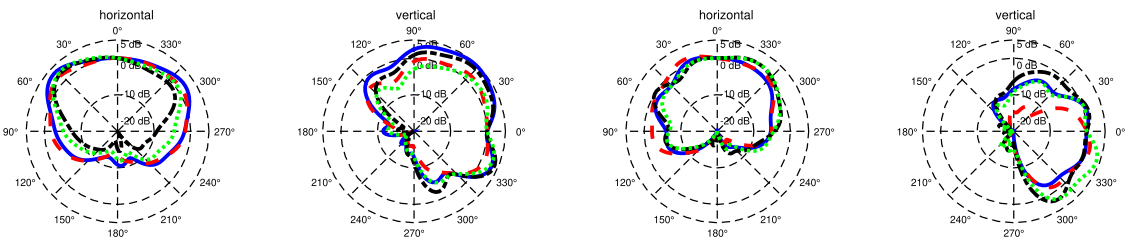
IV. DISCUSSION

A. Simulations

All three of the models used show that, in general, an increase of mouth opening in width and height leads to an increase in directionality. The transition frequency for the piston model can be identified by visual inspection of the normalized acoustic pressure maps. Nevertheless, the piston model does not predict sound radiation of a human properly. The main beam width below the transition frequency is too broad as the model does not account for the diffraction around the head, and no side lobes due to torso reflections can be predicted.

The normalized acoustic pressure maps of the spherical cap model do not reveal a clear transition from an omnidirectional to directional sound radiation because the diffraction of the sphere decreases the amplitudes of the sound radiated toward the side already at very low frequencies. The size of the mouth opening affects the main beam width and how sound is diffracted around the head, which can be seen for frequencies below 2 kHz for the angles between 90° and 180°. If the size of the mouth opening is increased, amplitudes at 180° increase and decrease around 135°. The effect of increased amplitudes toward the back below 1 kHz is often referred to as the acoustic bright spot (Hecht, 2002). The spherical cap model predicts the sound radiation of a human quite well, but experimental accuracy is still far out of reach as it becomes inaccurate at frequencies above roughly 4 kHz.

The results of the extended spherical cap model show that transverse propagation modes occur already around

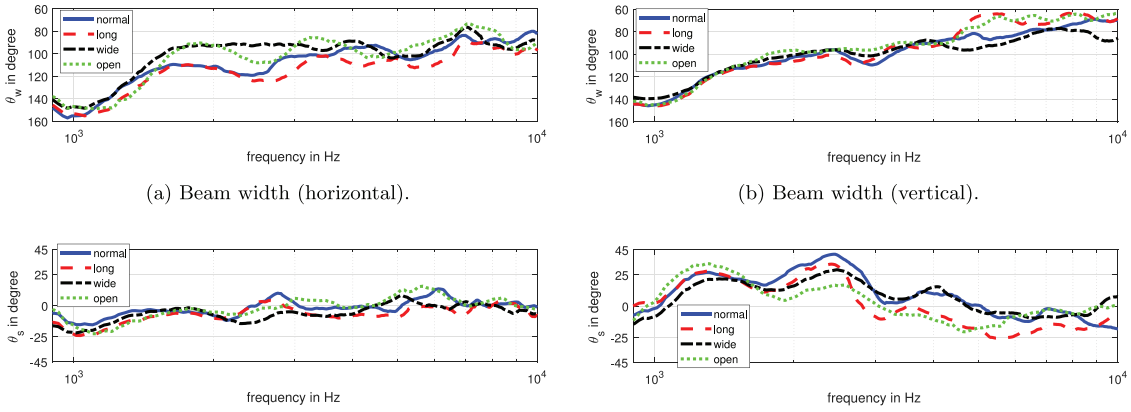


(a) 2400 Hz. (b) 5000 Hz.

FIG. 11. (Color online) Directivity patterns measured on the classical singer for the horizontal plane (left) and vertical plane (right) at (a) 2400 Hz and (b) 5000 Hz with the four mouth openings investigated (see Fig. 2): *normal* (solid blue), *long* (dashed red), *wide* (dashed-dotted black), and *open* (dotted green).

https://doi.org/10.1121/10.0001736

JASA



(a) Beam width (horizontal). (b) Beam width (vertical).

(c) Main direction (horizontal). (d) Main direction (vertical).

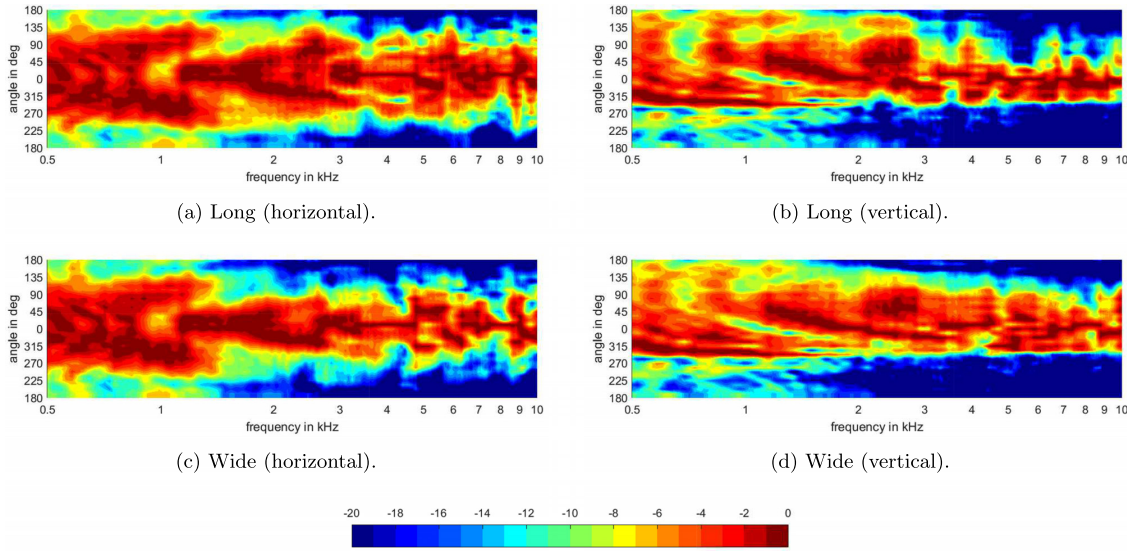
FIG. 12. (Color online) (a), (b) The beam width θ_w and [(c), (d)] the main direction θ_s of the energy vector are shown for the measurements performed on the classical singer with four different mouth openings in the horizontal and vertical planes.

3.6 kHz for the simulated vocal tract models. The transverse propagation modes create complex directivity patterns with strong local maxima and minima dependent on which mouth configuration is used. For larger mouth openings, the number of radiating transverse propagation modes increases significantly and, therefore, the complexity of the radiation pattern also increases significantly. The results show that the transverse propagation modes affect the main beam direction in the vertical plane increasing the downward deflection if the mouth opening is large (*long* and *open* configuration).

However, none of the three presented models take into account the effect of the torso. Another influence on sound radiation that is not included is the effect of the lips, which can be quite large dependent on the phoneme, as discussed in Yoshinaga *et al.* (2018).

B. Torso influence (HATS)

The directivity patterns measured on the HATS without torso in Fig. 6(a) agree very well with the simulation results for the spherical cap model in Fig. 4(a). Deviations occur above 2 kHz, where the measurement shows a broadening of the main beam width at several frequencies, which is not visible in the simulation. However, the head of the HATS has been only approximated by a sphere in the simulations. The on-axis frequency response reveals notches at the corresponding frequencies of the broadenings. These deviations



(a) Long (horizontal). (b) Long (vertical).

(c) Wide (horizontal). (d) Wide (vertical).

FIG. 13. (Color online) Normalized acoustic pressure maps (semitone-octave smoothed) as a function of the angular position and frequency measured on the classical singer for the *long* mouth configuration in (a) and (b) and for the *wide* mouth configuration in (c) and (d) in the horizontal and vertical planes, respectively.

JASA
https://doi.org/10.1121/10.0001736

are most likely edge diffraction effects of the head. Such diffraction effects can be partly simulated (Svensson, 2017; Vanderkooy, 1991) but have been omitted in the current study due to the complexity of the geometry of the HATS. The comparison of DI values in Fig. 7 shows a maximum deviation of around 1.6 dB between simulation and measurement. The negative DI values in Fig. 7 also confirm the limitation of the index as an optimal indicator for the characterization of speech or singing directivity.

Most striking as a result is the influence of the torso on the directivity for frequencies around 1 kHz. The torso diffraction is visible as large side lobes, which also provoke a decrease of amplitude of about 5 dB on-axis at 1.1 kHz. This explains a similar observation reported for singer directivity measurements in Katz and D'Alessandro (2007, Fig. 6).

C. Comparison of test subjects and HATS

The most obvious finding to emerge from the measurement data is that, dependent on the body dimensions, a decrease of amplitude toward the front around 1 kHz occurs for both test subjects and the HATS. As the test subject's dimensions increase, the decrease of the amplitude toward the front and, therefore, the first valley for the HDI, VDI, and FBR values is shifted toward a lower frequency (Fig. 8). This can be attributed to the fact that sound is diffracted by the torso, which we discuss in Sec. IV B. The torso diffraction provokes distinct side lobes at and above 1 kHz (Fig. 6).

By comparing polar patterns at 1.1 kHz, 800 Hz, and 950 Hz for the HATS, the amateur singer, and the classical singer, we can observe a similarity of voice directivity. The similarity of these polar patterns underlines the link between torso dimensions and torso diffraction and its effect on voice directivity in the frequency range around 1 kHz. As the frequency increases, and especially above 3.6 kHz, this similarity will most likely decrease as parameters such as the morphology of the vocal tract overtake the influence of the torso diffraction. Nevertheless, it is obvious that diffraction of the head also plays a role above 1 kHz.

D. Influence of the mouth configuration on radiation

We see from the measurement data that differences in sound radiation occur if the mouth configuration is changed to a larger extent. It is interesting to note that the results for the classical singer clearly show asymmetries along the vertical and horizontal planes. These changes of voice directivity and the asymmetries occur in both planes and were expected from the simulation results. All of the presented metrics reveal distinct differences between conditions. The change of mouth configuration provokes an increase or decrease of the main beam width and a shift of the main direction, which is seen in the analysis with the energy vector (Fig. 12). Two strong effects occur if the mouth is opened widely. A decrease of the main beam width is seen in the frequency region from 1.5 to 3 kHz for the *wide* and *open* mouth in the horizontal plane. This is also visible as an increase of the FBR values in the same frequency region. Furthermore, a change of main direction toward the floor is provoked at higher frequencies (>3 kHz) on the order of 20° by lowering the jaw (*long* and *open* conditions), which is seen for θ_s of the energy vector in the vertical plane and the UDR values. This downward focusing occurs even though we recognized a slight upward inclination of the head for the *long* and *open* conditions (cf. Table I). The downward focusing effect is most likely explained by the influence of transverse propagation modes, which can be seen in the results for the *long* and *open* conditions in the vertical plane in Fig. 5.

The acoustic pressure maps show highly complex patterns at higher frequencies (>3 kHz). The deviations from the simulations are most likely explained by the facts that (i) no MRI data were available to model the vocal tract of the classical singer precisely, (ii) the head is only approximated by a sphere, and (iii) the torso is not taken into account.

Still, the results for the singers agree with those of Kocon and Monson (2018), Blandin *et al.* (2018), and Katz and D'Alessandro (2007), which show that different mouth configurations (vowel and singing technique) provoke changes in the voice directivity. This study has been able to demonstrate that this effect even occurs for the German vowel /a/ if different mouth openings are used. Although the change of mouth opening introduces a deviation in a phonetic sense, it was used in this study as it is common in singing.

V. CONCLUSION

As very little was found in the literature on the question on how well simple models approximate human voice directivity, directivity models with different levels of complexity were compared with measurements of singers and a HATS. The simulations are also used to better understand detailed singing voice directivity measurement data. We tried to predict sound radiation for the classical singer with simulations, including the transverse propagation modes. We adapted the mouth opening dimensions in the simulations according to the extracted mouth opening height and width from the video used by the singer during the measurements. Prior studies that have noted deviations in voice directivity between test subjects have not investigated the specific effects that occur due to differences of torso dimensions and mouth openings for one vowel.

All our simulation models predict a higher directivity if the size of the mouth opening is increased. The extended spherical cap model predicts a strong influence on the voice directivity of the transverse propagation modes from 3.6 kHz on. In accordance with expectations, the study did not show that the simulation results fully reassemble the measurements and vice versa. However, the simulation results allow one to identify the combined effects seen in the voice directivity measurements of the singers.

It cannot be omitted that a detailed cause-and-effect analysis of voice directivity is quite difficult due to the many influences like posture, vowel tract geometry, and

https://doi.org/10.1121/10.0001736 JASA

subject differences. We minimized these effects in our study. The results show that a larger mouth opening reduces the main beam width by a magnitude of roughly 20° within the frequency range of 1–4 kHz and 5–9 kHz. This effect can be explained by a higher directivity due to a larger mouth opening, which also reduces diffraction at the torso. A second prominent effect is that a larger mouth opening provokes a shift of the main direction toward the floor for frequencies above 3 kHz.

With our analysis of voice directivity by simulations and measurements for different mouth openings, we give closer insight on singing voice directivity. The results indicate that sound can be focused to some extent toward the front by altering the mouth opening. We also show how the energy vector can be used as a more intuitive tool to analyze complex radiation patterns compared to the commonly used metrics.

Although the current study is based on a small number of participants, our findings agree well with results in the literature, which does allow us to generalize our findings to some extent. However, investigations on voice directivity in speech (Frank and Brandner, 2019) show that large changes in beam width are necessary to perceive an effect as a listener, which indicates that the observed effects may have little perceptual relevance. Therefore, the specific mouth configuration using in singing may be more directed toward increasing vocal production efficiency than on acting on the directionality.

ACKNOWLEDGMENTS

This work is partly supported by the project Augmented Practice-Room (1023), which is funded by the local government of Styria, and a part of this research was funded by the German Research Foundation (DFG), Grant No. BI 1639/7-1.

[1]See supplementary material at https://doi.org/10.1121/10.0001736 for the normalized acoustic pressure maps for the classical and amateur singers for all mouth opening configurations (SupMat.pdf). Furthermore, we include the results of the energy vector for the amateur singer and the measured mouth openings.

Aalto, D., Aaltonen, O., Happonen, R.-P., Jääsaari, P., Kivelä, A., Kuortti, J., Luukinen, J.-M., Malinen, J., Murtola, T., Parkkola, R., Saunavaara, J., Soukka, T., and Vainio, M. (**2014**). "Large scale data acquisition of simultaneous MRI and speech," Appl. Acoust. **83**, 64–75.

Abe, O. (**2019**). "Sound radiation of singing voices," Ph.D. thesis, Universität Hamburg, Hamburg, available at http://ediss.sub.uni-hamburg.de/volltexte/2019/9954 (Last viewed 13 May 2020).

Arnela, M., Dabbaghchian, S., Blandin, R., Guasch, O., Engwall, O., Van Hirtum, A., and Pelorson, X. (**2016**). "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," J. Acoust. Soc. Am. **140**(3), 1707–1718.

Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Hirtum, A. V., and Laval, X. (**2015**). "Effects of higher order propagation modes in vocal tract like geometries," J. Acoust. Soc. Am. **137**(2), 832–843.

Blandin, R., and Brandner, M. (**2019**). "Influence of the vocal tract on voice directivity," in *2019 Proceedings of the 23rd International Congress on Acoustics—ICA*, Deutsche Gesellschaft für Akustik e.V., pp. 1795–1801.

Blandin, R., Hirtum, A., Pelorson, X., and Laboissière, R. (**2016**). "Influence of higher order acoustical propagation modes on variable section waveguide directivity: Application to vowel [a]," Acta Acust. Acust. **102**, 918–929.

Blandin, R., Hirtum, A. V., Pelorson, X., and Laboissière, R. (**2018**). "The effect on vowel directivity patterns of higher order propagation modes," J. Sound Vib. **432**, 621–632.

Brandner, M., Frank, M., and Rudrich, D. (**2018**). "Dirpat—Database and viewer of 2D/3D directivity patterns of sound sources and receivers," in *Audio Engineering Society Convention 144*, available at http://www.aes.org/e-lib/browse.cfm?elib=19538 (Last viewed 17 July 2020).

Cabrera, D. (**2004**). "Vocal directivity measurements of eight opera singers," in *International Congress on Acoustics*, Vol. 1, pp. 505–506.

Cabrera, D., Davis, P. J., and Connolly, A. (**2011**). "Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment," J. Voice **25**(6), E291–E303.

Chu, W. T., and Warnock, A. C. (**2002**). "Detailed directivity of sound fields around human talkers," Technical Report No. RR-104, National Research Council of Canada, Institute for Research in Construction.

Farina, A. (**2000**). "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, available at http://www.aes.org/e-lib/browse.cfm?elib=10211 (Last viewed 26 May 2020).

Flanagan, J. L. (**1960**). "Analog measurements of sound radiation from the mouth," J. Acoust. Soc. Am. **32**(12), 1613–1620.

Frank, M. (**2013**). "Phantom sources using multiple loudspeakers in the horizontal plane," Ph.D. thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, available at http://phaidra.kug.ac.at/o:7008 (Last viewed 27 August 2020).

Frank, M., and Brandner, M. (**2019**). "Perceptual evaluation of spatial resolution in directivity patterns," in *Fortschritte der Akustik (DAGA 2019)* (*Progress in Acoustics*), edited by S. Spors and F. Wurm (Rostock, Deutschland).

Gerzon, M. A. (**1992**). "General metatheory of auditory localisation," in *Audio Engineering Society Convention 92*, available at http://www.aes.org/e-lib/browse.cfm?elib=6827 (Last viewed 8 March 2019).

Hecht, E. (**2002**). *Optics* (Addison-Wesley, Reading, MA).

Katz, B., and D'Alessandro, C. (**2007**). "Directivity measurements of the singing voice," in *19th International Congress on Acoustics*, Madrid, pp. 45–50.

Kob, M., and Jers, H. (**1999**). "Directivity measurement of a singer," J. Acoust. Soc. Am. **105**(2), 1003.

Kocon, P., and Monson, B. B. (**2018**). "Horizontal directivity patterns differ between vowels extracted from running speech," J. Acoust. Soc. Am. **144**(1), EL7–EL12.

Marshall, A. H., and Meyer, J. (**1985**). "The directivity and auditory impressions of singers," Acta Acust. Acust. **58**, 130–140.

Nair, A., Nair, G., and Reishofer, G. (**2016**). "The low mandible maneuver and its resonential implications for elite singers," J. Voice **30**(1), 128.e13–128.e32.

Savkar, S. (**1975**). "Radiation of cylindrical duct acoustic modes with flow mismatch," J. Sound Vib. **42**(3), 363–386.

Snakowska, A., Idczak, H., and Bogusz, B. (**1996**). "Modal analysis of the acoustic field radiated from an unflanged cylindrical duct—Theory and measurement," Acta Acust. Acust. **82**, 201–206.

Svensson, P. (**2017**). "Edge diffraction toolbox for MATLAB," available at https://github.com/upsvensson/Edge-diffraction-Matlab-toolbox (Last viewed 25 September 2019).

Tylka, J. G., Sridhar, R., and Choueiri, E. (**2015**). "A database of loudspeaker polar radiation measurements," in *Audio Engineering Society Convention 139*, available at http://www.aes.org/e-lib/browse.cfm?elib=17906 (Last viewed 13 March 2018).

Vanderkooy, J. (**1991**). "A simple theory of cabinet edge diffraction," J. Audio Eng. Soc. **39**(12), 923–933.

Wendt, F., Zotter, F., Frank, M., and Höldrich, R. (**2017**). "Auditory distance control using a variable-directivity loudspeaker," Appl. Sci. **7**, 666.

Xue, S. A., and Hao, J. G. (**2006**). "Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry," J. Voice **20**(3), 391–400.

Yoshinaga, T., Van Hirtum, A., Nozaki, K., and Wada, S. (**2018**). "Influence of the lip horn on acoustic pressure distribution pattern of sibilant /s/," Acta Acust. Acust. **104**(1), 145–152.

Zotter, F. (**2009**). "Analysis and synthesis of sound-radiation with spherical arrays," Ph.D. thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, available at http://phaidra.kug.ac.at/o:5619 (Last viewed 27 August 2020).

Zotter, F., and Frank, M. (**2019**). *Ambisonics. A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer, Berlin).

Zwicker, E., and Zollner, M. (**1993**). *Elektroakustik* (*Electroacoustics*), 3rd ed. (Springer, Berlin), Chap. 2, pp. 91–92.

2.6 Influence of Speech Sound Spectrum on the Computation of Octave Band Directivity Patterns

This work was published as:

R. Blandin, B. Monson, and **M. Brandner**. Influence of Speech Sound Spectrum on the Computation of Octave Band Directivity Patterns. *Proceedings of the FA2020 Conference*, 2027-2033. 2020. doi:10.48465/fa.2020.0446.

The idea and concept of this article were outlined by the first author. The first author wrote the original draft of the manuscript with periodical contributions from the second author and me. The revision was done with help from the second author and me.

INFLUENCE OF SPEECH SOUND SPECTRUM ON THE COMPUTATION OF OCTAVE BAND DIRECTIVITY PATTERNS

Rémi Blandin[1] **Brian Monson**[2] **Manuel Brandner**[3]
[1] Institute of Acoustics and Speech Communication , TU Dresden, Germany
[2] Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign, USA
[3] Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria
remi.blandin@tu-dresden.de

ABSTRACT

Speech radiation patterns exhibit angle-dependent variations of the amplitude and spectrum of the radiated sound. Speech directivity is gaining interest for the rendering of speech in three dimensional environments (real or virtual), but it is also related to more fundamental research questions, such as speech intelligibility in the presence of competing speech (the cocktail party problem). Speech directivity is most often quantified by octave-band analysis of speech signals recorded simultaneously with microphone arrays surrounding a talker in an anechoic environment. Due to the variability of the physical mechanisms of speech production, the radiation patterns differ between different speech sounds. However, a part of the observed variability may be due to the band analysis process itself, which is influenced by the spectral differences between the different speech sounds. In order to investigate to what extent and in which frequency range this variability is actually due to differences in directionality, directivity patterns are computed in narrower frequency bands with constant width. The details revealed by this higher spectral resolution also allow one to identify the expected influence of the dimensions of the subjects, the mouth opening and the contribution of the nasal cavity to the sound radiation. Octave band directivity patterns are computed from the high spectral resolution directivity patterns and compared with the commonly computed octave band long term averaged spectra directivity patterns.

1. INTRODUCTION

Speech directivity has been measured by several authors using multiple microphones placed at equidistant and regularly spaced positions around a talker (see as an example [1–4]). For most of these studies the sound levels recorded at the different angular positions have been measured in octave or third octave bands. In some cases, the analysis has targeted specific vowels and consonants, revealing directionality differences [1, 4–6].

The smooth polar diagrams presented in these studies can be misleading by suggesting that the directionality evolves progressively from one averaged pattern to the other as frequency increases. In fact, there is a lack of knowledge concerning the detailed features of human speech and singing directivity. Theoretical investigations of the effect of transverse propagation modes predict that, at least beyond 4-5 kHz, substantial variations in the directionality and shape of the directivity patterns can occur within small frequency intervals [7, 8]. These predictions have recently been confirmed by comparing simulations accounting for transverse propagation modes and the directivity patterns measured on two utterances of the vowel /a/ sung by a classical singer [9]. On the other hand, one can also expect that the mouth is not always the only source of sound radiation. For example, simultaneous radiation from the nose and the mouth might generate interference patterns at some frequencies.

Considering that directivity patterns can have significant variations within a relatively small frequency interval, this information is lost in an octave band pattern, which is essentially the average of multiple potentially different directivity patterns. Thus the overall octave band patterns are more heavily weighted by the patterns corresponding to the highest spectral amplitudes within the octave band. As a consequence, the observed differences in directivity between different speech sounds may be due to differences in the spectrum. On the other hand, differences in directivity may be missed simply because less energy is present at some frequencies within the octave band.

In order to investigate the details of speech directivity and the potential influence of the spectrum on octave band directivity patterns, directivity patterns of subjects recorded with a 13 microphone array in an anechoic room were computed in short time windows with 10.78-Hz spectral resolution. We examined the vowels /a/, /e/, /i/, /o/ and /u/. Octave band directivity patterns were computed following two methods:

- averaging the amplitude of the spectrum within the octave band,
- averaging normalized 10.78-Hz-wide directivity patterns for frequencies within the octave band.

The first method is similar to those employed in most previous studies, whereas the second method isolates the directivity phenomenon from the spectrum of the sound radiated.

2. METHOD

The data analyzed were sentences pronounced by 15 subjects (8 female and 7 male) recorded at a 44.1-kHz sampling rate with 13 microphones spaced at 15° intervals and equally distant from the head of the subjects in the horizontal plane, from 0° (directly in front of the talker) to 180° (directly behind the talker). Details of the acquisition of the data have been previously published [4]. The segmentation of the vowels /a/, /e/, /i/, /o/ and /u/ performed for a previous analysis of these data [6] was used here.

For the present study, the recordings from each microphone were sliced in windows of 2048 samples overlapping by 90%. The spectrum was computed using a discrete Fourier transform and zero padding (2048 additional samples) so that the frequency resolution was 10.78 Hz. These parameters were the same for both methods detailed hereafter.

2.1 Averaging of the amplitude of the spectrum

Before averaging, unwanted noise was removed from the data. A pure tone at 17.6 kHz was removed. The microphone at 165° had a higher background noise than the other microphones, which created artefacts when the signal to noise ratio was low. Consequently, the data for this microphone were excluded from the averaging process from 5 kHz on. Because this microphone is located behind the head of the subject where the radiated amplitude is low at high frequencies, it did not substantially affect the analysis of the directivity patterns.

An average spectrum for each vowel was generated by averaging across each windows for all the utterances of each vowel by all subjects (see as an example Figs. 2a and 2c for the 500Hz octave band interval). The octave band directivity patterns for each vowel were then computed by calculating the average amplitude of the averaged spectrum over the frequency interval corresponding to each octave band (see Fig. 3). The directivity patterns were then normalized by the maximum amplitude across all angles.

2.2 Averaging of the directivity patterns

Directivity patterns were computed for each frequency and each time window by subtracting the maximal amplitude over the angular positions from the amplitude of the other positions. Thus, the spectral level was normalized at each frequency, allowing each directivity pattern at each frequency to be equally weighted during the averaging process.

However, before computing directivity patterns, it was necessary to define a frequency dependent noise threshold. This is necessary because computing directivity patterns from data with too poor signal to noise ratio would lead to wrong patterns. For example, if a highly directional pattern is observed with a poor signal to noise ratio, only the highest amplitudes would emerge from the noise.

A 9s background noise recording was used to compute a noise threshold for each microphone. Spectra of overlapping windows were computed in the same way as for the other data. The median of the background noise amplitude was computed for each frequency and each microphone in order to get a smooth approximation of the background noise profile. The obtained curve was shifted up in level so as to exceed the maxima of all windows from the noise signal. The 17.6 kHz peak was added to the threshold. Any data having an amplitude lower than this threshold curve were excluded from the analysis.

Directivity patterns were computed only if at least 3 microphones registered amplitude higher than the noise threshold. To visualize the directivity patterns, they were represented in color scale as a function of the frequency and the angular position (see Figs. 1 and 2). This representation is referred to as a directivity map hereafter.

The obtained directivity maps were averaged over time windows for individual utterances of each vowel for each subject (see Figs 1a and 1b), for all utterances of all vowels for each subject (see Figs 1c and 1d), and for all utterances of all subjects for each vowel (see Figs 2b and 2d for the frequency interval corresponding to the 500Hz octave band). Octave band directivity patterns were computed averaging the directivity maps over octave bands (see Fig. 4).

3. RESULTS

3.1 Directivity maps

Figs. 1a and 1b show the directivity maps (calculated by averaging directivity patterns) for single utterances of the vowels /a/ and /i/ pronounced by the same female subject. The evolution of the directivity patterns with increasing frequency show abrupt transitions to different patterns and directionality within small frequency intervals (on the order of 100Hz). For example, in Fig. 1a there is a sudden appearance of a pattern with lower directionality near 8kHz. More complexity can generally be found toward high frequencies. It is noteworthy that the complex variations of the directivity patterns with increasing frequency are also variable in time. They change from utterance to utterance and even possibly within a single utterance (data not shown).

Directivity maps averaged on all utterances of individual subjects are presented in Figs. 1c and 1d, for a female and a male subject respectively. One can see that the complexity is substantially reduced compared to the unique utterances (Figs. 1a and 1b). A global increase of directionality can be seen up to about 10 kHz. Above 10 kHz, the directionality appears to slightly decrease.

All directivity maps in Fig. 1 reveal a similar pattern at frequencies below 6 kHz, consisting of 3 or 4 lobes diverging toward the side as frequency increases. Slight changes in the number of lobes and their angular position are observed between different subjects and between different utterances within a given subject. In some utterances, other patterns featuring substantial changes within small frequency intervals appear superimposed to this lobe pattern, as seen in Fig. 1b in the 0-2 kHz interval.

In Figs. 2a and 2c the average spectrum is represented

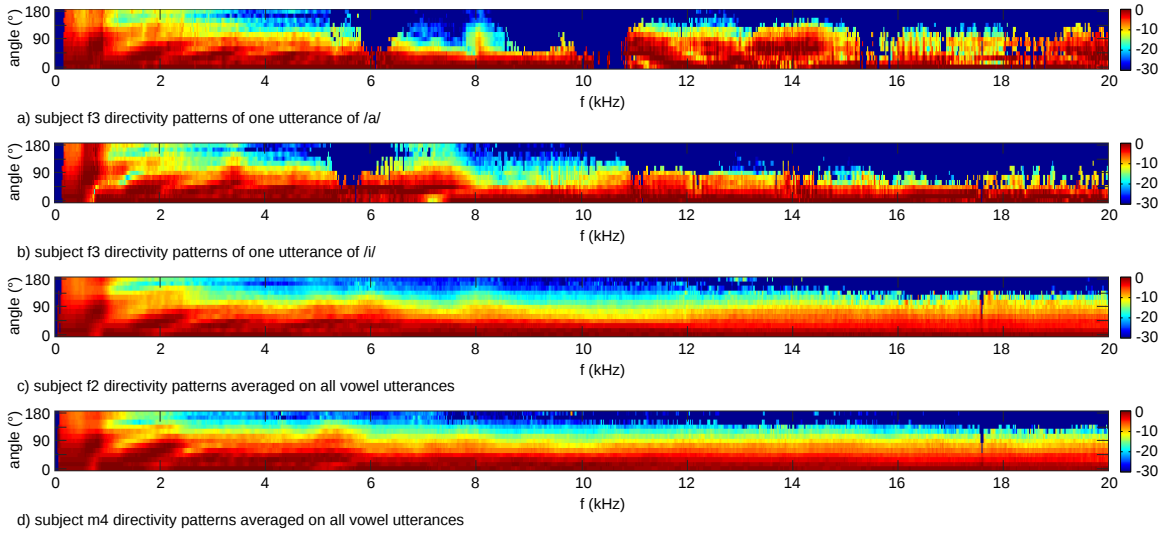


Figure 1. Directivity maps (calculated by averaging directivity patterns) obtained with a 10.78 Hz discretization represented in color scale as a function of the frequency and the angular position. (a) a single utterance of the vowel /a/ by a female subject, (b) a single utterance of the vowel /i/ by the same female subject, (c) directivity patterns averaged on all utterances of all vowels of one female subject and (d) directivity patterns averaged on all utterances of all vowels of one male subject. The color indicates the sound level relative to the maximal level over the 13 positions for each frequency. The deep blue color corresponds to data under a noise threshold.

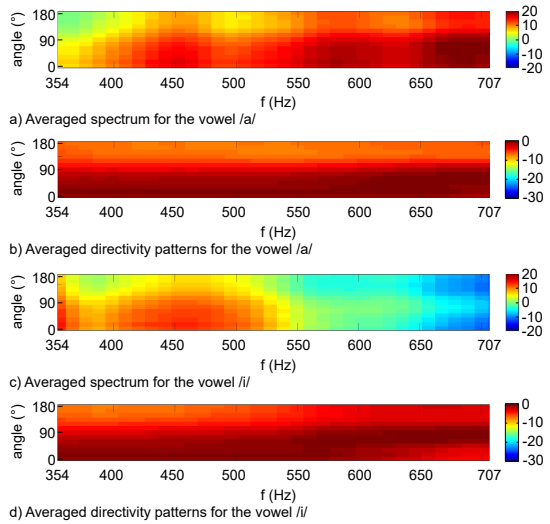


Figure 2. Amplitude of the spectrum and directivity patterns averaged on several utterances of the vowel /a/ and /i/ pronounced by 15 female and male subjects in the frequency interval corresponding to the 500Hz octave band. (a) averaged amplitude spectrum for the vowel /a/, (b) averaged directivity patterns for the vowel /a/, (c) averaged amplitude spectrum for the vowel /i/, and (d) averaged directivity patterns for the vowel /i/.

in color scale over the frequency range corresponding to the 500Hz octave band (354-707Hz) for the vowels /a/ and /i/. The directivity maps corresponding to the same frequency range and vowels are presented in Figs. 2b and 2d.

The averaged spectrum amplitude and the directivity patterns of the vowels /a/ and /i/ are presented in Fig. 2 for the frequency range corresponding to the 500Hz octave band (354-707Hz). The acoustic energy appears more evenly distributed for the vowel /a/ (Fig. 2a) than for the vowel /i/ (Fig. 2b). The energy of the vowel /i/ is mainly present in the first half of the frequency band (354-550Hz), and substantially lower amplitudes are found between 650Hz and 707Hz. The directivity patterns of the vowel /a/ (Fig. 2b) are rather similar all over the 500Hz frequency band, with a slight sideward shift of the maximum of amplitude from 600Hz on. The directivity patterns of /i/ (Fig. 2d) show more variations: a more pronounced lobe oriented toward 90° can be seen from 600Hz on.

3.2 Octave band directivity patterns

The directivity patterns obtained from averaged spectra and averaged directivity patterns are presented in Figs. 3 and 4, respectively. 0° corresponds to the front and 180° to the back of the subjects. Both methods generate globally similar patterns: the shape of the patterns is very similar and the directionality increases with the frequency of the octave bands, except for the 16kHz band.

However, there are substantial differences in the 8kHz and 16kHz band: the patterns obtained from averaged spectra (Fig. 3) are more directional. Another noticeable difference is the greater variability of the patterns across

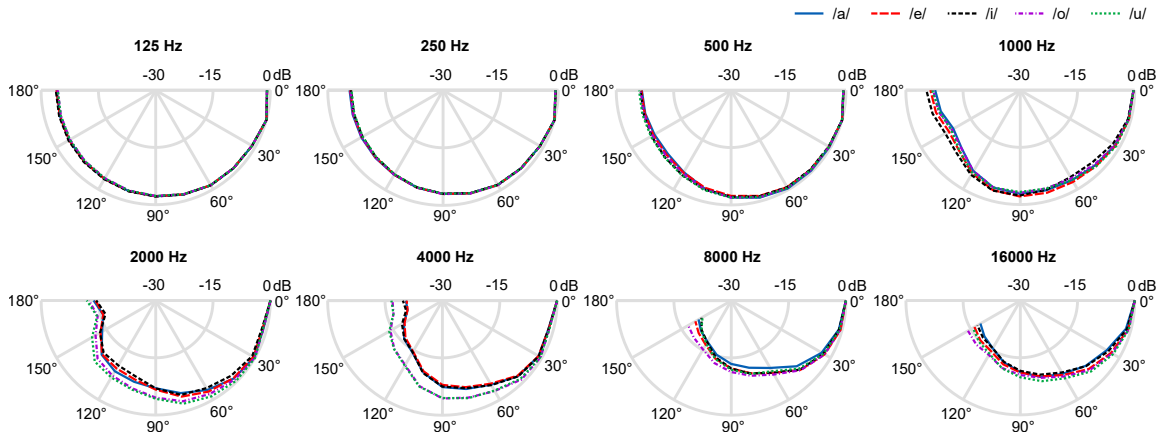


Figure 3. Directivity patterns obtained averaging the sound amplitude over octave bands and over the utterances of 15 male and female subjects for the vowels /a/, /e/, /i/, /o/ and /u/. The radius of the curves indicates sound level relative to the maximal level over the 13 positions.

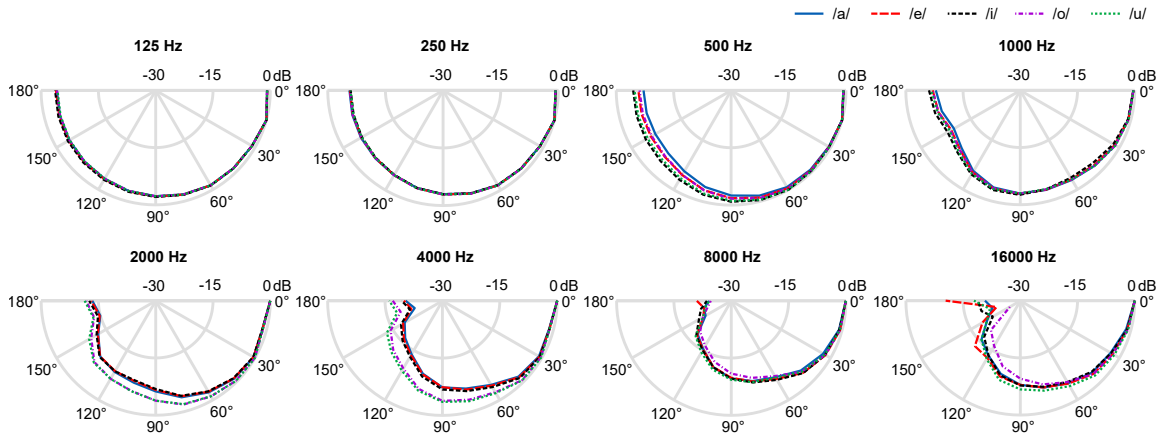


Figure 4. Directivity patterns obtained averaging the directivity patterns computed every 10.78 Hz over octave bands and over the utterances of 15 male and female subjects for the vowels /a/, /e/, /i/, /o/ and /u/. The radius of the curves indicates sound level relative to the maximal level over the 13 positions.

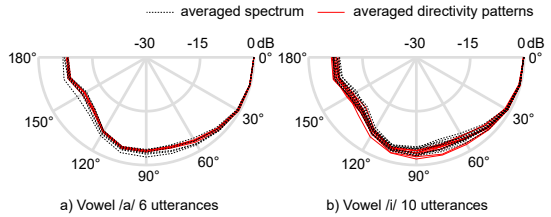


Figure 5. Directivity patterns computed for multiple individual utterances of the same male subject from averaged spectrum (dashed black lines) and from averaged directivity patterns (full red lines) for (a) the vowel /a/ and (b) the vowel /i/.

vowels in the 500Hz band for the patterns obtained from averaged directivity patterns (Fig. 4). This variability appears to be greater than that observed in the 1kHz band. In the 2kHz band, the vowels are more clearly separated in two groups in the octave band patterns from averaged directivity patterns.

For both octave band patterns computation methods, in the 125Hz and 250Hz band exactly the same patterns are obtained for all the vowels. Differences appear in the 500Hz band, in which the vowels are, in order of increasing directionality, /i/, /u/, /e/, /o/ and /a/. Less differences are observed in the 1kHz band, but one can note that the vowel /i/ has a slightly different pattern from the others, with a slightly lower amplitude in the 30°-90° region and the highest amplitude in the 90°-180° region. In the 2kHz and 4kHz bands the vowels are separated in two groups: /a/, /e/ and /i/ more directional and /u/ and /o/ less directional. In the 8 kHz and 16kHz bands there is no more separation in two groups and less differences between the vowels than in the 2kHz and 4kHz bands.

Fig. 5 show the patterns obtained for multiple utterances of the vowels /a/ and /i/ pronounced by the same male subject. In the case of the vowel /a/, less utterance to utterance variations are observed when using averaged directivity patterns. In the case of the vowel /i/, the variability of the patterns is similar for both methods.

4. DISCUSSION

4.1 Comparison of two methods of computation of octave band directivity patterns

The main difference between octave band directivity patterns computed from averaged spectra (Fig. 3) and from averaged 10.78Hz directivity patterns (Fig. 4) is the increased directionality obtained with the first method in the 8kHz and 16kHz octave bands. This can be explained by the expected effect of unequal weighting of patterns at individual frequencies, as pointed out in the introduction. Because these bands cover the most extended frequency ranges, they are the most likely to be influenced by this problem. In the 16kHz band, more acoustic energy is present in the lower frequencies of the band at which the directivity patterns are more directional than in the higher frequencies (see Figs. 1c and 1d). Thus, with these patterns being over-represented, the resulting octave band pattern is more directional than the one obtained from averaged 10.78Hz directivity patterns. In the case of the 8kHz band, the situation is reversed: the directionality increases with frequency in the band and there is less acoustic energy at lower frequencies of the band. The less directional patterns being less weighted, the overall pattern is also more directional than the one obtained from averaged 10.78Hz directivity patterns.

Another difference is observed in the 500Hz octave band: the directivity patterns obtained for each vowel by averaging the spectrum are almost identical, whereas substantial differences across vowels are found when averaging the 10.78Hz directivity patterns. This can be explained by an uneven distribution of the acoustic energy over the frequency band. For the vowel /i/, the less directional pattern compared to the other vowels is due to the presence of a pronounced lobe orientated toward 90° in the upper part of the band (see Fig. 2d). Its orientation, which differs from the other patterns of the band, makes the averaged overall pattern less directional. However, the formant structure of the vowel /i/ is such that there is less energy in the upper portion of the band (see Fig. 2c). Thus, the weight of this 90° lobe is small in the averaging process, and this difference of directivity compared to the other vowels is underestimated. On the other hand, the acoustic energy of the vowel /a/ is more evenly distributed in this band (see Fig. 2a), but the 90° lobe is less pronounced than for the vowel /i/ (see Fig. 2b). Thus, averaging the spectrum to compute the octave band directivity pattern tends to underestimate the differences in directionality of the vowels in the 500Hz octave band.

At the scale of single utterances, greater variability in the octave band directivity patterns obtained from averaged spectra (see Fig. 5a) can also be attributed to variations in the distribution of the acoustic energy over the bands. However, the acoustic energy distribution also affects the averaging of 10.78Hz directivity pattern, because there are frequencies at which the amplitude is smaller than the noise threshold. On the other hand, the directivity patterns themselves vary from utterance to utterance, especially at high frequencies. Similar variability can be observed with this method with utterances of other speech sounds (see Fig. 5b). Thus, using multiple utterances to fill acoustic energy gaps when averaging out the variability of directivity patterns is important.

The use of a noise threshold allowed us to obtain information from the two back microphones (165° and 180°), which receive the smallest amplitudes. However, due to the smaller amount of data, the averaged patterns obtained in this angular region may be less accurate. This may be why the shape of the pattern obtained for /e/ departs from the patterns obtained for the other vowels in the 16kHz octave band (see Fig. 4).

4.2 Potential mechanisms inducing vowel directivity pattern variations

The complex variations of the directivity patterns obtained from averaged 10.78Hz directivity patterns (Figs. 1, 2 and 4) can be explained for high frequencies (from 4-5 kHz on) by the propagation of transverse propagation modes inside the vocal tract. In fact, it has been shown theoretically and experimentally that this phenomenon can generate complex variation of the directivity patterns with frequency [8, 9].

However, transverse propagation modes do not explain the abrupt transitions observed at low frequencies, observed in Fig. 1b in the 0-2 kHz interval. This may be due to the nasalisation of the vowels which sometimes occurs when the communication with the nasal cavity is open. In such a case, sound would be radiated simultaneously by the nose and the mouth and produce interference patterns which might be responsible for these abrupt transitions. This needs to be confirmed by proper modelling of the phenomenon.

Some key parameters in determining the directionality of speech are the dimensions of the mouth opening, specifically its width. From the plane piston radiation model, one expects that the wider is the mouth, the more directional the sound is in the horizontal plane [10, 11]. Directionality is expected to increase with the frequency.

However, this tendency is not always observed on the averaged directivity maps of the different subjects (see Figs. 1c and 1d): the directionality slightly decreases from about 10kHz on. On the other hand, the differences of mouth opening would be expected to have a stronger impact at high frequencies. Again, this is not always observed, as less differences between vowels are observed in the 8 kHz and 16 kHz octave bands than in the 2 kHz and 4 kHz octave bands. The transverse propagation modes may be the explanation of this disagreement with the simple plane piston model. Their more frequent occurrence at high frequencies, their tendency to generate less directional patterns and their variability from utterance to utterance would result in an average decrease of directionality. However, this needs to be confirmed by proper modelling. It should be noted that this is an average tendency and that at the scale of single utterances the patterns can be very different from the averaged patterns, as illustrated in Figs. 1a and 1b. The relevance of this variability over time and frequency of the high frequency directivity patterns for the perception of human speech and singing is an open question.

On the other hand, the plane piston model explains very well the division of the directivity patterns in two groups in the 2 kHz and 4kHz octave bands. In fact, the same two groups, /a/, /e/ and /i/ which are more directional and /o/ and /u/ which are less directional, are found when sorting the vowels by mouth width. The mouth widths measured by Fromkin [12] for different vowels show that /a/, /e/ and /i/ corresponds to similar mouth width (about 40mm) which are greater than the ones of /o/ (about 20mm) and /u/ (about 15mm). One can even notice that the vowel /o/ is slightly more directional than the vowel /u/ in the 4 kHz octave band in Fig. 4, which is in agreement with the dimensions provided by Fromkin.

However, the plane piston model fails again to predict the variations of directionality in the 500Hz and 1kHz octave bands:

- more differences are observed in the 500Hz band than in the 1kHz band, whereas one would expect the opposite,
- the differences of directionality are no more correlated to the mouth width, /i/ being the less directional whereas it has a larger width than /o/ and /u/.

In fact the plane piston model would not predict strong variations of directionality with the width, and finding significant variation in the 500Hz octave band is surprising. The reason of this disagreement is not clearly understood. A possible explanation could be that the nasalisation of the vowels plays a role in the directivity of the vowels in these octave bands. Nasalisation may induce complex variations of directivity patterns with strong minima, such as the ones observed in Fig 1b. This could result in the 90° lobe observed in the upper part of the 500Hz band (see Fig. 2) when averaging over several utterances. Some vowels may tend to be more nasalized than others and, thus have different patterns in these octave bands. Other alternative explanations could be the differences in protrusion, or radiation from other parts of the head such as the cheeks or the larynx. A proper modelling of the potentially implied phenomena as well as a tracking of the nasalization is needed to clarify these various hypothesis.

The lobe pattern between 0 and 4 kHz is most likely due to the diffraction by the torso. In fact, similar lobes are predicted for the head related transfer function, which are very close to the reciprocal situation of speech production [13]. The inter- and intra- subject variations of this pattern can be explained by differences in the dimensions, shape of the head and the torso of the subject as well as their posture.

5. CONCLUSION

The computation of directivity patterns at multiple frequencies with a small frequency spacing (10.78Hz) reveal that speech directivity has complex variations on small frequency scales (on the order of 100Hz). These variations are the consequences of different phenomena implied in speech and singing radiation. More physical modelling is required to identify them. The process of averaging spectrum to build octave band directivity patterns appears to overestimate directionality in the 8kHz and 16kHz bands, and to underestimate the variability related to vowels in the 500Hz. This is the expected consequence of an uneven weighting of directivity patterns resulting from the uneven distribution of the acoustic energy over the speech spectrum. Thus, the process of directly averaging high frequency resolution directivity patterns appears to be more accurate and reliable. However, it requires the definition

of noise thresholds. In this purpose, it is important to perform a recording of background noise free from perturbation from the subjects (breathing) for the measurement of directivity.

6. REFERENCES

[1] A. Marshall and J. Meyer, "The directivity and auditory impressions of singers," *Acta Acustica united with Acustica*, vol. 58, no. 3, pp. 130–140, 1985.

[2] W. Chu and A. Warnock, "Detailed directivity of sound fields around human talkers," 2002.

[3] D. Cabrera, P. Davis, and A. Connolly, "Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment," *Journal of Voice*, vol. 25, no. 6, pp. e291–e303, 2011.

[4] B. Monson, E. Hunter, and B. Story, "Horizontal directivity of low-and high-frequency energy in speech and singing," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 433–441, 2012.

[5] B. Katz and C. d'Alessandro, "Directivity measurements of the singing voice," 2007.

[6] P. Kocon and B. B. Monson, "Horizontal directivity patterns differ between vowels extracted from running speech," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. EL7–EL12, 2018.

[7] R. Blandin, A. Van Hirtum, X. Pelorson, and R. Laboissière, "Influence of higher order acoustical propagation modes on variable section waveguide directivity: Application to vowel [α]," *Acta Acustica united with Acustica*, vol. 102, no. 5, pp. 918–929, 2016.

[8] R. Blandin, A. Van Hirtum, X. Pelorson, and R. Laboissière, "The effect on vowel directivity patterns of higher order propagation modes," *Journal of Sound and Vibration*, vol. 432, pp. 621–632, 2018.

[9] R. Blandin and M. Brandner, "Influence of the vocal tract on voice directivity," 2019.

[10] J. Flanagan, "Analog measurements of sound radiation from the mouth," *The Journal of the Acoustical Society of America*, vol. 32, no. 12, pp. 1613–1620, 1960.

[11] J. Huopaniemi, K. Kettunen, and J. Rahkonen, "Measurement and modeling techniques for directional sound radiation from the mouth," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*, pp. 183–186, IEEE, 1999.

[12] V. Fromkin, "Lip positions in american english vowels," *Language and speech*, vol. 7, no. 4, pp. 215–225, 1964.

[13] V. Algazi, R. Duda, R. Duraiswami, N. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2053–2064, 2002.

2.7 Horizontal and Vertical Voice Directivity Characteristics of Sung Vowels in Classical Singing

This work was published as:

M. Brandner, M. Frank, and A. Sontacchi. Horizontal and Vertical Voice Directivity Characteristics of Sung Vowels in Classical Singing. *MDPI Acoustics*. 4:849–866, 2022. doi:10.3390/acoustics4040051.

The idea and concept of this article were outlined by me, the first author, with help from the second and the third author. I wrote the original draft of the manuscript. The revision and editing was done by me with help from the second author. I did all of the programming, visualizations, measurements and prepared the data for publication.

acoustics

MDPI

Article

Horizontal and Vertical Voice Directivity Characteristics of Sung Vowels in Classical Singing

Manuel Brandner *, Matthias Frank and Alois Sontacchi

Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, 8010 Graz, Austria
* Correspondence: brandner@iem.at

Abstract: Singing voice directivity for five sustained German vowels /a:/, /e:/, /i:/, /o:/, /u:/ over a wide pitch range was investigated using a multichannel microphone array with high spatial resolution along the horizontal and vertical axes. A newly created dataset allows to examine voice directivity in classical singing with high resolution in angle and frequency. Three voice production modes (phonation modes) modal, breathy, and pressed that could affect the used mouth opening and voice directivity were investigated. We present detailed results for singing voice directivity and introduce metrics to discuss the differences of complex voice directivity patterns of the whole data in a more compact form. Differences were found between vowels, pitch, and gender (voice types with corresponding vocal range). Differences between the vowels /a:, e:, i:/ and /o:, u:/ and pitch can be addressed by simplified metrics up to about d2/D5/587 Hz, but we found that voice directivity generally depends strongly on pitch. Minor differences were found between voice production modes and found to be more pronounced for female singers. Voice directivity differs at low pitch between vowels with front vowels being most directional. We found that which of the front vowels is most directional depends on the evaluated pitch. This seems to be related to the complex radiation pattern of the human voice, which involves a large inter-subjective variability strongly influenced by the shape of the torso, head, and mouth. All recorded classical sung vowels at high pitches exhibit similar high directionality.

Keywords: singing voice directivity; classical singing; voice directivity metrics; directivity index; musical acoustics

check for updates

Citation: Brandner, M.; Frank, M.; Sontacchi, A. Horizontal and Vertical Voice Directivity Characteristics of Sung Vowels in Classical Singing. *Acoustics* **2022**, *4*, 849–866. https://doi.org/10.3390/acoustics4040051

Academic Editors: Muhammad Naveed Aman, Anwar Ali and Asif Iqbal

Received: 27 July 2022
Revised: 24 September 2022
Accepted: 27 September 2022
Published: 1 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Studies on voice directivity have been investigating several aspects and generally agree about the factors that influence voice directivity the most. Voice directivity is determined by the morphology of a person, posture, vocal tract shape and the effective mouth opening. Investigations on voice directivity have been undertaken for verification, auralization, or performance analysis of the human voice. Previous studies have included measurements of sound radiation from artificial mouths, human talkers [1–7], and singers [8–15] with various approaches in regard to spoken or sung content and microphone array setups.

Voice directivitiy is reported to be affected by different vowels which are exhibiting unique radiation characteristics according to [5,7] and showed highest directionality for the vowel /a/ in [11] or /ɑ/ in [6], followed by /e/, /i/, /o/, and /u/, although for singers in [8] the vowel /e/ is reported to be most directional. In general, the effective difference is shown to be rather subtle for speech but expected to be larger in classical singing due to larger mouth openings. However, in singing the inter-subjective variability is reported to be substantially large [12] which could diminish a clear effect of differences in voice directivity for vowels for a specific subject group.

Voice directivity characteristics can be calculated from measured sound pressure levels at fixed distances from a defined center position along a circle or sphere around a person. The levels can be acquired by time domain or frequency domain analysis.

Acoustics **2022**, *4* 850

These characteristics inform about how sound is radiated from a singer or talker. The most common metric, and usually used for loudspeaker or antenna characteristics, is the directivity index [16,17]. Recently, we also introduced the beam width and direction of the energy vector metric [15,18] and discussed its usefulness as a descriptor for voice directivity, since the focus of this metric is not front-centered compared to the directivity index. This is especially useful along the vertical axis. The metrics can be computed from long-term averaged spectra (LTAS) or from levels computed from bandpassed signals for each spoken or sung phoneme or phrase. It is also possible to calculate and discuss voice directivity characteristics from impulse responses of sung glissandi (vocal sweeps) [7,15] or from the LTAS of a sung glissandi directly [9]. This approach needs a reference microphone in front of the mouth and a training phase for the subjects to keep their vocal tract configuration constant during the recording of the glissandi.

An important factor for the analysis of voice directivity is the spectral distribution of a single phoneme. Therefore, the mode of voice production in terms of level [5,11] and vocal fold vibration (phonation mode at the vocal folds [19]) influence the energy in different frequency bands and hereby the effective measurable radiated sound. The phonation mode defines the degree of impulsiveness during vocal fold closure of the source signal at the vocal folds and introduces a certain spectral tilt [20]. General differences in phonation modes are described by the terms breathy, modal, and pressed phonation, although other phonation modes especially in pathological voices exist, which are not addressed within this study. The phonation mode implies an effect on the effective sound radiated from the mouth due to spectral changes. However, for classical singing, there is a secondary factor worthwhile investigating, namely the directivity of the voice. The phonation mode may introduce a change in the mouth opening used by the singer which has not yet been investigated in previous studies.

We previously demonstrated that mouth opening and body size have an effect on the voice directivity characteristics of singers [15]. These findings imply that singers of different voice types (e.g. soprano, tenor, etc.) and therefore with different vocal ranges should exhibit different voice directivity characteristics for the same vowel identity. In previous studies, gender differences have been addressed but the results for an effect of gender on voice directivity disagree [5]. For singing, it would make more sense to discuss this in terms of vocal range rather than gender, as we expect the effective mouth opening to increase with pitch for each voice type. However, in this study, the vocal range criterion again separates singers by gender as well.

In the current study we investigate the following:

- the effects of vocal range (voice type) on voice directivity characteristics,
- the effects of mouth opening in regard to pitch and vowel,
- vowel specific radiation characteristics for classical singers in the horizontal and vertical plane,
- the influence of the phonation mode on voice directivity characteristics.

This contribution explains the employed measurement system, the newly generated dataset, and the methods used to process the measurement data. We present results from the acoustic data as polar patterns, and simple broadband or frequency-dependent metrics. Furthermore, we present tracking and video data that allows us to investigate the influence of the mouth opening more rigorously. The findings of the work are interesting for the fields of performance analysis, musical applications, audio recording, virtual and augmented reality systems.

2. Materials and Methods

2.1. Measurement System

A measurement system in an anechoic chamber for the determination of singing voice directivity was set up using the double circle microphone array (DCMA) [21], ten optical tracking sensors and a video camera in order to measure the mouth opening and center position of the singer. The video camera allows to validate the measurement results from the

tracking system, to show exemplary mouth openings, and opens the possibility to calculate the mouth opening from video directly. The measurement software was implemented in Pure Data (freely available under http://puredata.info/, accessed on 26 July 2022).

2.2. Room Conditions

Measurements were carried out in a sound treated measurement room with absorptive material on the walls and floor at the Institute of Electronic Music and Acoustics. The mean room reverberation between 400 Hz and 1 kHz is below 75 ms and above 1 kHz below 50 ms. The volume of the room is approximately 50 m^3 with a floor area of 22.50 m^2.

2.3. Double Circle Microphone Array

The used microphone array has a radius of 1 m and consists of two circular rings, one placed in the horizontal plane and one in the vertical plane [21]. Each ring holds up to 32 microphones (Omnidirectional pattern, NTI M2230, Schaan, Liechtenstein) resulting in an angular spacing of 11.25° and a total number of 62 microphones (see Figure 1). In addition, a reference microphone (NTI M2230) is used, which is located at the exact center of the microphone array, but not considered in the current calculations.

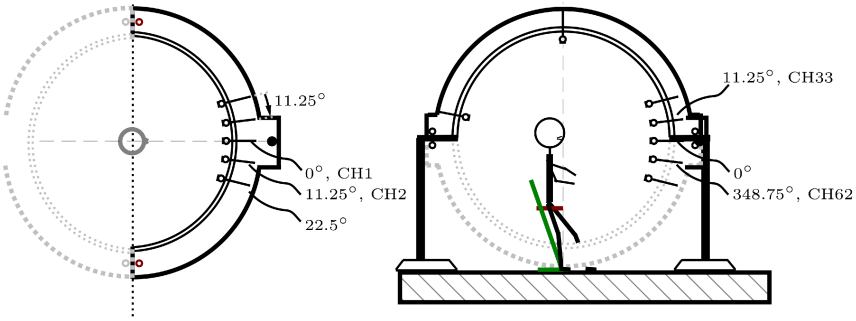


Figure 1. (**Left**): Top view on the microphone array (horizontal plane). (**Right**): Side view on the microphone array (vertical plane). The singer is seated on a chair of adjustable height. Microphone positions and angular spacing are indicated.

2.4. Mouth Tracking

The mouth opening and absolute position of the singer inside the microphone array is captured by a tracking system (Optitrack (https://www.optitrack.com/)). The tracking system uses ten cameras (Flex 13), six of them are positioned in the corners of the ceiling in the anechoic chamber and four closer in front of the singer (in 1 m distance). The closer cameras increase the localization accuracy for the mouth tracking. The absolute position is used to prevent large positioning errors during measurements. The conductor of the measurements and the participant have both a visual feedback of the current position, which indicates larger deviations than 5 cm from the center with a warning. If such a warning should occur, the measurement is repeated. The distance of 5 cm has been found to be acceptable [15] and the largest expected error for the sound pressure level at a single microphone position is 0.72 dB. The tracking data is then used to calculate the mouth opening area. Therefore, the tracking data of the single facial markers around the mouth are put in order by calculating the convex hull for each time frame. The resulting polygons are used to calculate the average area that gives the effective mouth opening area. Note that the positioning of the facial markers as close as possible to the lips leads to a minor individual offset, which is expected to be constant for each singer during the measurements. However, as we are interested in the relative change due to pitch and vowel, we do not compensate for this constant offset in the further analysis.

2.5. Augmented Acoustics

For the acoustic analysis, dry signals are measured in an anechoic environment. However, in singing, room acoustics support the voice, which is a necessity in a longer recording session. Therefore, we use an augmented acoustic system with zero latency [22,23] that only gives the singer natural room acoustics via transparent headphones [24] while creating no reverberation on the microphone signals. The augmented acoustic system is fed by the microphone in front of the singer and employs a static, however frequency-dependent directivity to excite the virtual room. The virtual room simulates a shoe-box-like concert hall with a size of roughly 30 m $\times$ 24 m $\times$ 20 m and a reverberation time of 2.2 s. Typical reverberation times of concert halls are in the range between 1.5 s and 3 s [25,26].

2.6. Dataset

For the purpose of this study we created a newly dataset of 5 sustained German vowels /a:/, /e:/, /i:/, /o:/, /u:/ sung by 4 male singers (3 tenors, 1 baritone) over the pitch range on a whole-tone scale from H/B2/123 Hz to a^1/A4/440 Hz, except for the baritone only up to e^1/E4/330 Hz, and from six female singers (3 soprano and 3 mezzo-soprano) from a/A3/220 Hz to a^2/A5/880 Hz. All singers were trained classical singers except for the baritone (jazz), who said to have the ability to mimic the classical singing technique due to his teaching experience at the music conservatory. The average age was 29.6 years; the youngest 24 years old, the oldest 34. The classical trained singers were 4 graduate (at the end of their current master studies), 5 post-graduate students (with one master's degree ore more), and 1 undergraduate (bachelor's degree). Six of them were also teaching. The singers were asked to sing the vowels, starting on the consonant /m/ and sustaining the vowel for 2 seconds. The vowels were repeated three times each with different provoked voice phonation modes (modal, breathy, and pressed), which gives a total number of 2145 audio samples. The singers were asked to sing at a comfortable loudness level (mezzo-forte). All participants were well trained for the task due to their extensive practice during their classical vocal studies. The dataset is publicly available under https://phaidra.kug.ac.at/o:127031 (accessed on 26 September 2022).

2.7. Calculation of Directivity Characteristics

For the current study, a large dataset with multiple variables was created, which means that simplified metrics may be beneficial for discussing potential differences in directivity between vowels, gender, pitch, and phonation modes. For our simplified metrics for the whole data, we opted for a broadband approach. Directivity characteristics can be computed in the frequency domain or time domain. The general difference between these two calculation methods lies in the different consideration of frequency components. As mentioned above, the spectral components differ depending on the vowel identity and phonation mode and therefore influence the effectively radiated sound. In our investigation, we want to focus on the maximum separability of the data. Therefore, we compute directivity characteristics for a sung vowel at a single pitch (i) from frequency data calculated using Welch's method (averaged periodogram method) [27] to discuss general differences between the female and male singers (see Section 3.1), and (ii) from levels from frequency domain data extracted only at the harmonics (see Section 3.3). Averaging of metrics calculated from frequency domain data means a stronger weighting of spectral components at higher frequencies that usually exhibit lower sound pressure levels, whereas these components would have only a smaller influence on metrics calculated broadband in the time domain. Nevertheless, including the frequency components equally will allow a better discrimination between vowels, for example. The audio signals for the frequency analysis are segmented with a frame length of 93 ms and 50% overlap at a sampling frequency of 44.1 kHz. Then, the spectrum of each segment is calculated with a frequency resolution of approx. 10.7 Hz. The estimated averaged frequency responses are then third-octave smoothed. The segments extracted from the recordings for further analysis exclude the consonant at the beginning of each vowel. The resulting data allows to

investigate differences of voice directivity between sung vowels, which exhibit complex radiation patterns if analyzed in detail, with simplified metrics.

2.8. Directivity Patterns

Instead of neglecting a signal analysis focusing only on components of high sound pressure level, we present compact results from high-passed signals (1 kHz, 4th-order Butterworth) as polar patterns evaluated in both planes for female and male singers. The cutoff frequency of the high-pass filter is chosen according to the findings in Section 3.1. A quasi-continuous representation of arbitrary radiation directions for the azimuth angle ϕ can be rendered from given discrete measurement positions by applying the circular harmonic transform [28]. The polar patterns are displayed logarithmically and show a dynamic range of 25 dB in each plane.

2.9. Directivity Index

The directivity factor $\gamma_p(\omega) = \frac{P_{on-axis}}{P_{mean}}$, the most common metric in directivity analysis [16], in each plane is defined by the ratio of the on-axis power $P_{on-axis}$ to the average power P_{mean} of all sampling positions on the respective plane. The horizontal and vertical directivity index (HDI, VDI), evaluated at an angular frequency ω is defined in dB as follows:

$$DI(\omega) = 10 \log_{10}(\gamma_p(\omega)). \tag{1}$$

However, due to its definition the directivity index is front centric and has been shown to decrease strongly at frequencies around 550 to 1000 Hz for the human voice [7,15]. This decrease is dependent on the shape and size of the torso, head size, and vowel. The effect of reduced frontal radiated energy also occurs at odd multiples of the first strong valley in the directivity index when investigated over frequency [15]. However, in order to perform a comparative analysis with previous studies, we include the directivity index results in the current study.

2.10. Beam Width of the Energy Vector

The energy vector is commonly used in the context of 3D loudspeaker playback, but is as well useful in the description of the characteristics of any arbitrary sound source radiation [18] and avoids the directivity index problem of frontal fixation on a single measurement point. This is especially useful for the vertical plane, because previous studies report a more downward radiating voice at higher frequencies [8,15]. The energy vector $\boldsymbol{r_E}$ in Equation (2) can be utilized to describe the main radiating direction and its corresponding width of an acoustic source.

$$\boldsymbol{r_E}(\omega) = \frac{\sum_{i=1}^{L} |H(\omega, \phi_i)|^2 \boldsymbol{m_i}}{\sum_{i=1}^{L} |H(\omega, \phi_i)|^2}. \tag{2}$$

The frequency-dependent magnitudes $H(\omega, \phi_i)$ at the measurement angles ϕ_i are multiplied by the vectors $\boldsymbol{m_i} = [\cos(\phi_i), \sin(\phi_i)]^T$ of each measurement position i, $i = 1, 2, ..., L$ in each respective plane, and normalized by the sum of the energy, yielding a normalization of the vector between the limits 0 (omni-directional) to 1 (maximum focus to one direction). As a non-front-centered metric in comparison to the directivity index we use the main beam width in each plane in Equation (3):

$$\theta_w = 2 \arccos \|\boldsymbol{r_E}\|. \tag{3}$$

The beam width θ_w will measure the broadness of the beam towards the direction of highest intensity (see Figure 2). In the case of two side-lobes with similar strength and a

decrease towards the front, the energy vector $\boldsymbol{r_E}$ will be still centered towards the front but exhibit a broader beam width. In the case of a single side-lobe being stronger than the other, the direction of $\boldsymbol{r_E}$ will change more towards the direction of the competing side-lobe dependent on its level.

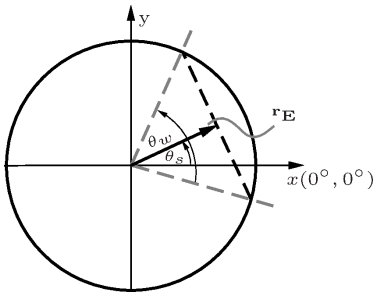


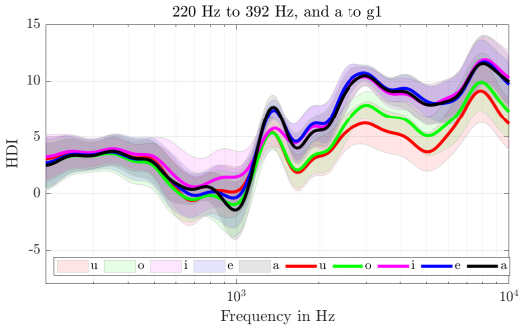
Figure 2. Schematic of the energy vector and its corresponding source angle θ_s and source width θ_w.

3. Results and Discussion

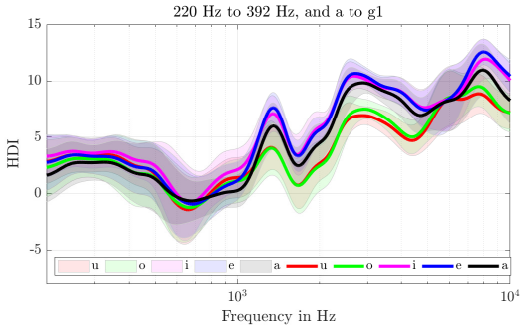
3.1. Effects of Vocal Range

The first analysis aims on the relationship of vocal range on voice directivity. Therefore, we present the results for HDIs and VDIs averaged over the pitches a/A3/220 Hz to g1/G4/392 Hz for the female and male singers separately. The singers have an overlapping pitch range from a/A3/220 Hz to a1/A4/440 Hz. We present the mean values and standard deviations of the HDIs averaged over the overlapping pitch range in Figure 3A,B for each vowel. The results agree well with the voice directivity study of 13 talkers in [7], but most strikingly the classical singers show an overall higher directivity above 2.5 kHz compared to the talkers. Furthermore, in the figures is shown that the first prominent decrease in HDIs starts around 650 Hz for female and male singers, but only the female singers exhibit a distinct decrease around 1 kHz. This could be linked to the influence of the torso size [15] in combination with the effective mouth opening used by the singers. The effective mouth opening differs for the same vocal range between the singer groups, which is also shown by the results presented in the next section. The larger used mouth openings of the male singers (cf. Section 3.2) explain the smaller differences between HDIs and VDIs for the front vowels /a:, e:, i:/, and back vowels /o:, u:/, compared to the female results. Another interesting aspect is that in the vertical plane, the vowels /e:, i:/ are more directional on average above 1.3 kHz. This could be due to stronger reflections of the upper body in regard to the used mouth opening for these vowels.

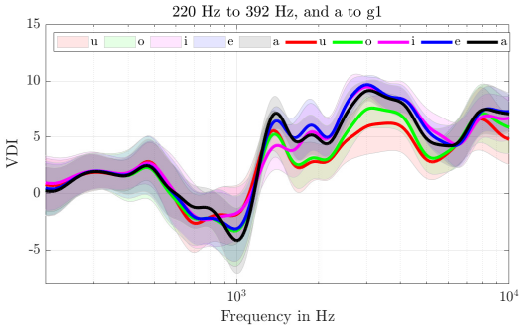
In addition and as alternative to the directivity index, we present the beam width metric, which does not exhibit extreme decreases due to its mathematical definition, but shows similar pronounced differences to discuss effects on voice directivity and can be found in Table A1. In Figure 4 we plot the directivity patterns at low (a/A3/220 Hz) and high (a1/A4/440 Hz)) pitch computed from high-pass signals at 1 kHz with a 4th-order Butterworth filter to exclude torso influences and include only frequency components with high sound pressure levels and to see if this already reduces the differences between the genders (voice range). Most striking in the figures is that the gender or voice range difference is still quite visible for the horizontal plane with more pronounced backward radiation for the female singers at both pitches. It is also shown that the male singers exhibit less variability between vowels, which was already seen by the metrics presented in Figure 3. The male singers exhibit broader patterns in the horizonal plane for /e:, i:/ compared to the female singers. In general we see large variability in the data and a dependence on pitch. Again, the differences in the data between female and male singers seems to be related to the mouth opening, which is discussed in the next section in more detail.



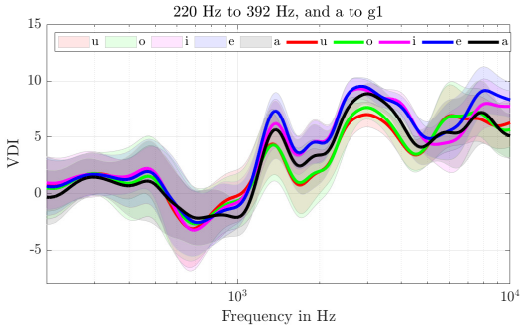
(**A**) Female singers (HDI).



(**B**) Male singers (HDI).



(**C**) Female singers (VDI).



(**D**) Male singers (VDI).

Figure 3. Mean and standard deviation for the horizontal and vertical directivity index in dB averaged over pitch (a/A3/220 Hz to g1/G4/392 Hz) for the female singers (**A**,**C**) and male singers (**B**,**D**) for all five vowels and all phonation modes.

3.2. Effects of Pitch on Mouth Opening

The measurement setup allows to track the effective mouth opening by using facial markers and a motion tracking system (The tracking of single markers proved to be problematic for some of the male singers, so the tracking data is omitted for the male singers and video extracts are presented instead.). The effective mouth opening is computed by calculating the convex hull [29] of the positions of the facial markers in a Cartesian coordinate system. Figure 5 shows that the average mouth opening areas for the vowels of the six female singers differ. The average difference between all vowels up to d^2/D5/587 Hz lies at 0.98 cm^2 with average differences between each vowel pairs: $\overline{\Delta}_{ae}$ = 0.95 cm^2, $\overline{\Delta}_{ei}$ = 1.59 cm^2, $\overline{\Delta}_{io}$ = 0.21 cm^2, and $\overline{\Delta}_{ou}$ = 1.58 cm^2. Above d^2/D5/587 Hz, a general increase in mouth opening is observed, rising to a maximum of about 15 cm^2. At the highest pitch a2/A5/880 Hz the mouth opening area decreases slightly for the front vowels /e/ and /i/. Figure 6 shows how differently the mouth opening is shaped for all vowels at low (c/C3/196 Hz) and high pitch (a1/A4/440 Hz) for one male singer and in Figure 7 for one female singer. Here, the expected difference due to the classical singing style between mouth openings for female and male singers (singers of different vocal range) is already apparent. The male singers already reach their highest pitch at a1/A4/440 Hz and use larger mouth openings for all vowels, while the female singers still use moderate mouth openings at the same pitch with visible differences shown in Figure 7.

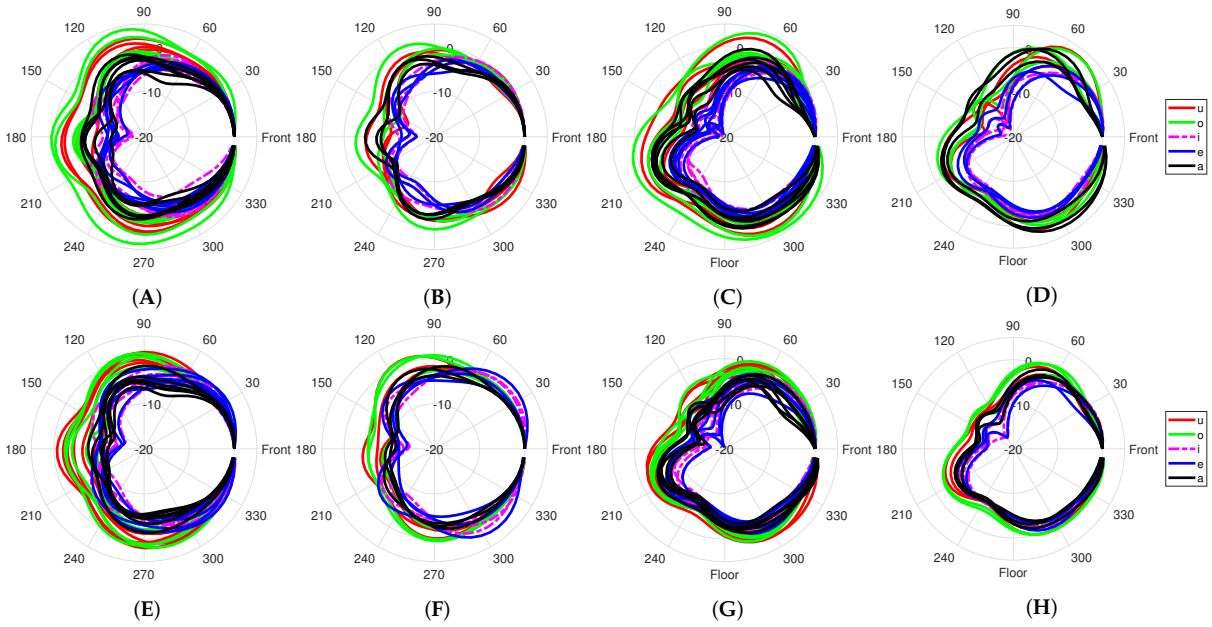


Figure 4. Directivity patterns for female and male singers in the horizontal and vertical plane at low (a/A3/220 Hz) and high (a1/A4/440 Hz) pitch for all five vowels of high-passed signals (1 kHz). The multiple lines of the same color represent the results for each singer for a specific vowel. The vowel data is averaged over all phonation modes. (**A**) Female singers (horizontal, a/A3/220 Hz), (**B**) Male singers (horizontal, a/A3/220 Hz), (**C**) Female singers (vertical, a/A3/220 Hz), (**D**) Male singers (vertical, a/A3/220 Hz), (**E**) Female singers (horizontal, a1/A4/440 Hz), (**F**) Male singers (horizontal, a1/A4/440 Hz), (**G**) Female singers (vertical, a1/A4/440 Hz), (**H**) Male singers (vertical, pitch a1/A4/440 Hz).

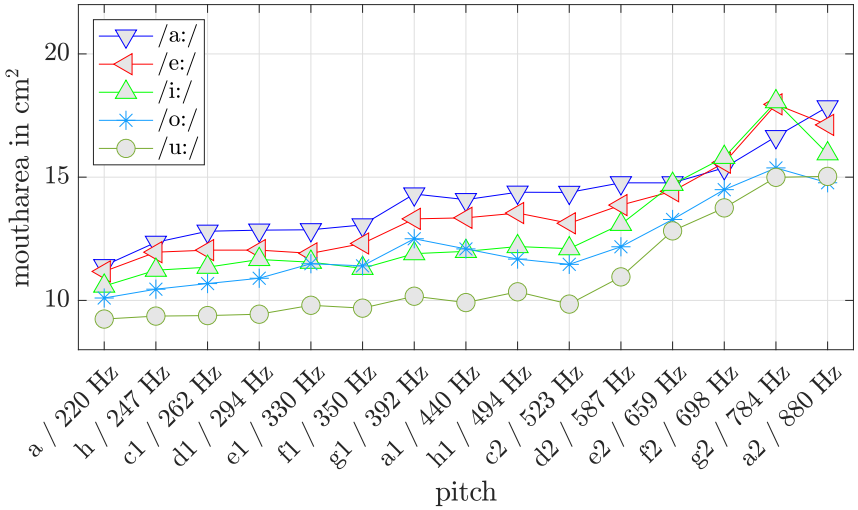


Figure 5. Mean effective mouth opening calculated from tracking data for the five vowels from six female singers.

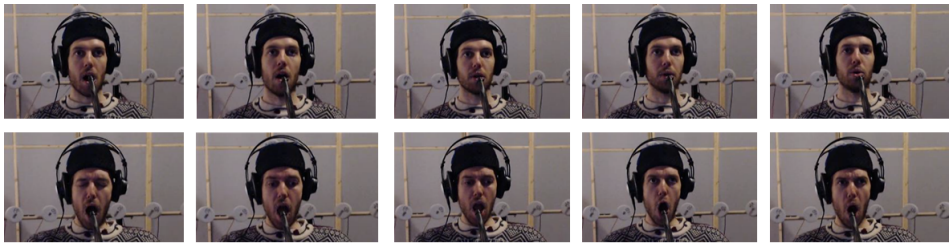


Figure 6. Exemplary mouth openings extracted from video for one male singer for the vowels /a:/, /e:/, /i:/, /o:/, /u:/ (left to right) at c/C3/131 Hz in the top row and a^2/A4/440 Hz in the bottom row.

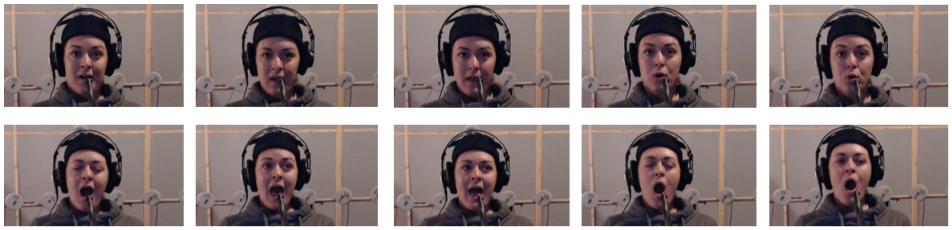


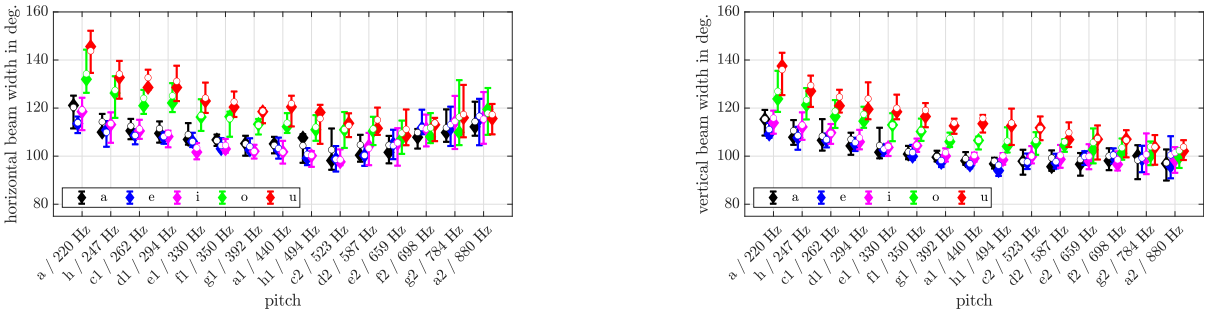
Figure 7. Exemplary mouth openings extracted from video for one female singer for the vowels /a:/, /e:/, /i:/, /o:/, /u:/ (left to right) at a^1/A4/440 Hz in the top row and a^2/A5/880 Hz in the bottom row.

3.3. Beam Width and Directivity Index for Female Singers

In this and the next section, we present detailed results for our simplified metrics calculated from levels extracted from frequency domain data only at the harmonics. A short discussion on the trade-off between calculation methods of simplified metrics can be found in Section 2.7. Figure 8 shows the means, medians, and IQRs (inter-quartile ranges) of the beam width in the horizontal and vertical plane. The data presents the vowel sequence /a:/, /e:/, /i:/, /o:/, and /u:/ over all pitches sung by the female singers (a/A3/220 Hz to a^2/A5/880 Hz). As each vowel is sung in three voice phonation modes, we get a total number of 18 measurements per pitch for the female singers (6 singers, 5 vowels, 3 voice phonation modes). As the data is paired (the singers sung multiple vowels at different phonation modes) and the distributions are not normally distributed (Lilliefors test, $p > 0.05$), we test the differences by using the Wilcoxon signed rank test [30] and use a Bonferroni-Holm correction [31] to account for the five groups. We underpin our findings by using the biserial correlation coefficient [32,33] (effect size) to measure the practical significance, whereas detailed results are provided in Table A1 for the horizontal and vertical beam width (As a general rule of thumb effect sizes are considered for an r = 0.10 as small, r = 0.30 as medium and r $\geqslant$ 0.50 as large effect sizes [34].).

We find differences at a significance level of $p < 0.05$ for the front vowels /a:/, /e:/, /i:/ and the back vowels /o:/, and /u:/ until pitch d^2/D5 for the beam width, except for the pair /i:/ and /o:/ (biserial correlation coefficient of 0.6 in both planes). In the horizontal and vertical plane, the beam width decreases over pitch (cf. Figure 8A,B). This can be explained by the fact, that with increasing pitch the overall human voice directivity is sampled differently by the harmonics (cf. Figure 3). Interestingly, the decrease ends at about d^2/D5 and indicates a migration of all vowels towards a similar directivity at high pitch. Most striking in the figures is the decrease of the horizontal beam widths at higher pitch. These results are likely to be related to two factors, namely (i) a reduction of the mouth width and more focus on lowering of the jaw of the singers at high pitch and (ii) a more downward focused voice at higher pitch, which has already been discussed in [15]. In contrast, the vertical beam width increases steadily with increasing pitch. Closer

inspection of the figures shows that, depending on the pitch, different front vowels are most directional. This result supports the various findings from previous studies about which vowel is most directional.

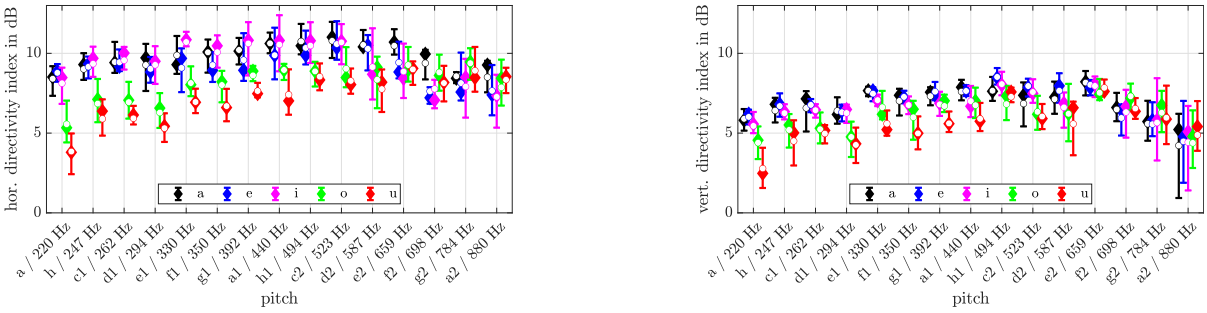


(**A**) Horizontal beam width.

(**B**) Vertical beam width.

Figure 8. Means (diamonds), medians (dots), IQRs (whiskers) of the horizontal and vertical beam width in degree over pitch for the female singers for the five vowels and all phonation modes.

Figure 9 shows the voice directivity analyzed by the directivity index metric in the horizontal and vertical plane. We find differences ($p < 0.05$) for the front vowels and the back vowels below pitch c^2/C5/523 Hz (see detailed results in Table A2). In general, the directivity index of the different vowels becomes similar towards higher pitch and tends to decrease for pitches higher than d^2/D5/587 Hz with a stronger decrease in the vertical plane above f^2/F5/698 Hz (cf. Figure 9A,B). The general trend of the directivity index in both planes is very similar to the beam width, increasing towards higher pitch, and then decreasing slightly above d^2/D5/587 Hz. The data for the directivity index show less statistical and practical significance at higher pitches between vowels (cf. Table A2) compared to the beam width metric.



(**A**) Horizontal directivity index.

(**B**) Vertical directivity index.

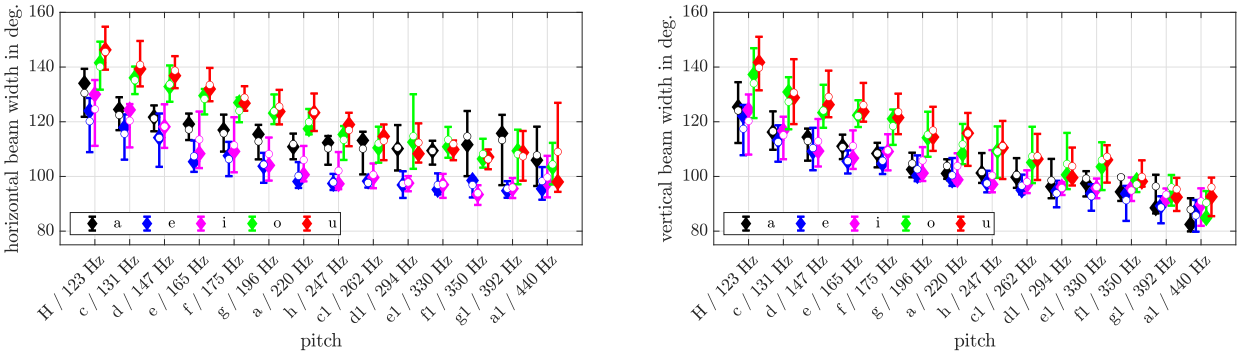
Figure 9. Means (diamonds), medians (dots), IQRs (whiskers) of the horizontal and vertical directivity index in dB over pitch for the female singers for the five vowels and all phonation modes.

3.4. Beam Width and Directivity Index for Male Singers

Figure 10 shows the means, medians, and IQRs of the beam width in the horizontal and vertical plane for the male singers. The presented data include all four male singers for the pitch range H/B2/123 Hz to e^1/E4/330 Hz and include the three tenors up to a^1/A4/440 Hz. Again, each vowel is sung in three voice phonation modes, giving us a total number of 12 measurements per pitch up to e^1/E4/330 Hz and 9 measurements per pitch above. Again, the distributions for the male singers are not normally distributed (Lilliefors test, $p > 0.05$) and the data is paired. Therefore, we test the differences with the Wilcoxon signed rank test and use a Bonferroni-Holm correction. The biserial correlation coefficient in detail is shown in Table A3.

We find differences at a significance level of ($p < 0.05$) for the beam width of the front vowels /a:/, /e:/, /i:/ and the back vowels /o:/, and /u:/ for the horizontal and vertical beam width above pitch d^1/D4/294 Hz (cf. Table A3). Similar as for the female singers, a steady decrease for all the vowels is shown with increasing pitch (cf. Figure 10A). This decrease ends in the horizontal plane around d^1/D4/294 Hz and remains around the same level for higher pitches in contrast to the slight decrease for the female singers (cf. Figure A1). In the horizontal plane, the beam width of the front vowels at a^1/A4/220 Hz (lowest pitch of the female singers) are about 100° to 120° and therefore a magnitude of 20° lower than for the female singers. This may be related to the larger mouth openings used by the male singers, which has been discussed in Section 3.2. In the vertical plane, the beam width of the front vowels decrease more steadily with increasing pitch compared to the results shown in the horizontal plane (cf. Figure 10B). This indicates a more prominent vertical mouth opening for the male singers with increasing pitch. Nevertheless, the results in Figure 3 in Section 3.1 show a difference between gender below 1 kHz, suggesting that this also has an effect on the metric and is a cause of the more even decrease in beam width for the male singers.

Furthermore, we find differences at a significance level of ($p < 0.05$) for the front vowels /a:, e:, i:/ and the back vowels /o:, u:/ up to pitch c1/C4/262 Hz for the horizontal and vertical directivity index (cf. Figure 11 and Table A4). Again, the statistical results attest the directivity index almost the same quality to distinguish between front and back vowels as the beam width metric.



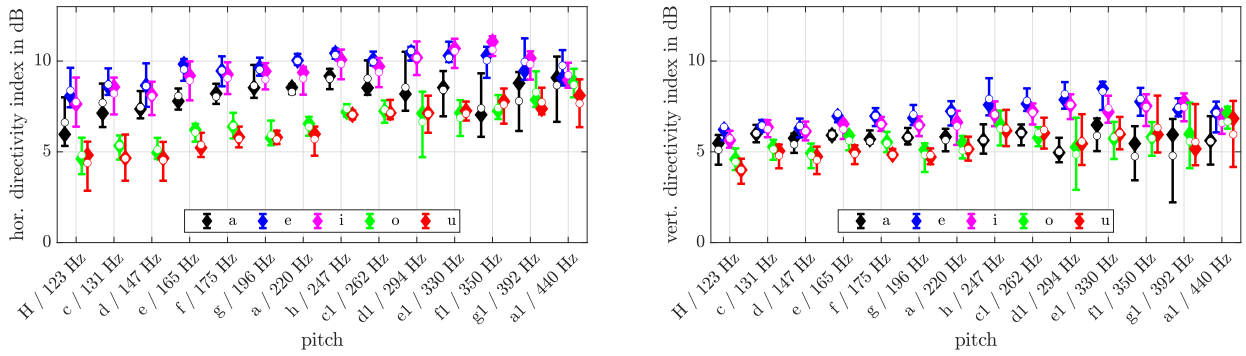
(**A**) Horizontal beam width.

(**B**) Vertical beam width.

Figure 10. Means (diamonds), medians (dots), IQRs (whiskers) of the horizontal and vertical beam width in degree over pitch for the male singers for the five vowels and all phonation modes.

3.5. Voice Directivity Characteristics and Phonation Mode

To investigate whether voice directivity characteristics are affected by phonation modes, we present results for the beam width for the female and male singers analyzed by their medians and IQRs over all pitches for each vowel and phonation mode (see Figure 12). The means of the medians of the female group for the horizontal plane in Figure 12A lie at 107° for the front vowels and at 118° for the back vowels. In the vertical plane, the beam width is slightly lower (101° for front vowels, 110° for back vowels) for all vowels (Figure 12C) with a trend to decrease from breathy to pressed phonation mode. This effect is larger when evaluated only at higher pitches, which can be also seen in the tracking data. Higher variance is shown for the back vowels /o:, u:/ and slightly higher for breathy compared to the other phonation modes. This can be explained by a tendency of the singers to prefer larger mouth openings for the back vowels over intelligible articulation. The medians for the male group are smaller for the vowels /e:, i:/ and higher for the vowels /a:, o:, u:/ compared to the female singers as shown in Figure 12B. Phonation modes for the male singers reveal no clear difference or trend for the beam width in the vertical plane (cf. Figure 12D), but overall show higher variances compared to the female singers.

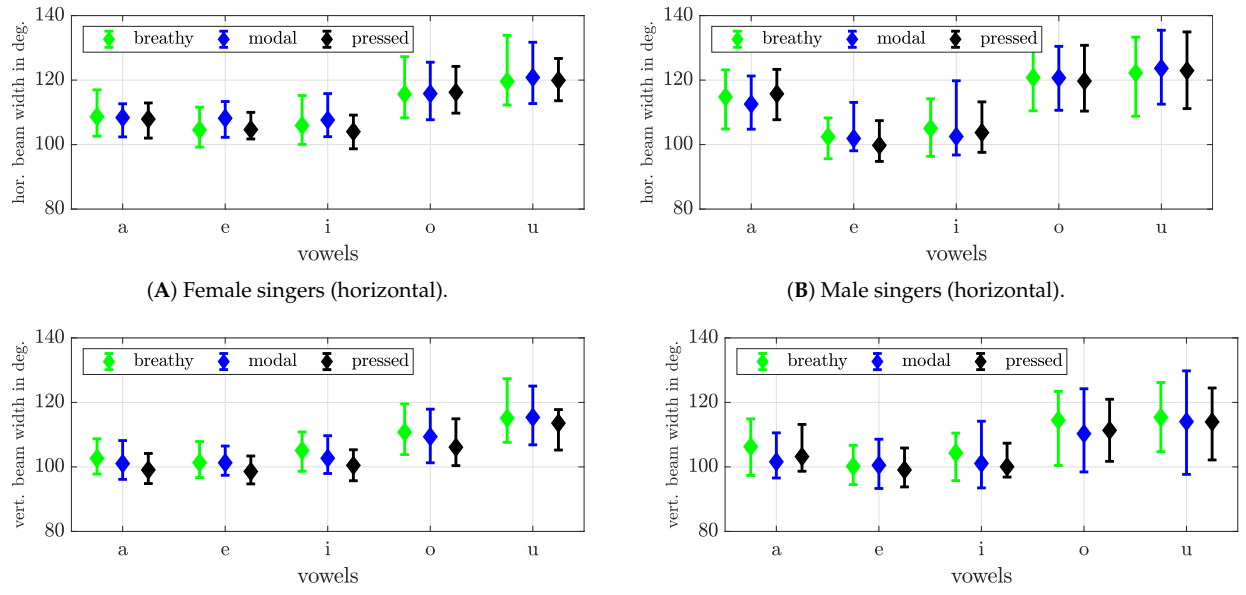


(**A**) Horizontal directivity index.

(**B**) Vertical directivity index.

Figure 11. Means (diamonds), medians (dots), IQRs (whiskers) of the horizontal and vertical directivity index in dB over pitch for the male singers for the five vowels and all phonation modes.

The averaged metrics over pitch indicate a slightly trend of increased mouth openings from breathy to pressed. This is also linked to an increased effort at the vocal folds (breathy, modal, pressed) which could be noticed in the audio and during the recording session. Furthermore, the general difference between front /a:, e:, i:/ and back vowels /o:, u:/ can be made visible by the metrics. Most striking is the result for the male group of lower beam widths in the horizontal plane for the vowel /e:, i:/ indicating broader mouth openings for the male singers than the female singers. However, the generalisability of this result seems to be rather limited due to the small sample size.



(**A**) Female singers (horizontal).

(**B**) Male singers (horizontal).

(**C**) Female singers (vertical).

(**D**) Male singers (vertical).

Figure 12. Medians and inter-quartile ranges for the beam width over all pitches in the horizontal (**A**,**B**) and vertical (**C**,**D**) plane for each vowel and phonation mode for the female and male singers.

4. Conclusions

The present study investigates voice directivity in classical singing and its dependence on vowel, pitch, and gender (vocal range). In general, we exhibit higher directivity for classical singers when compared to voice directivity in speech. The current data show

that mouth opening increases for classical singers of different vocal ranges (gender) and that this is also linked to pitch. We found subtle differences between the female and male singers (vocal ranges) at lower frequencies from 650 Hz to 1 kHz and minor differences at higher frequencies. The differences at higher frequencies can be explained by the different used mouth openings within the respective vocal range (larger for male singers), whereas at lower frequencies the differences are most likely linked to the size of the torso. The simplified metrics calculated from frequency data showed the capability of separating the front vowels /a:, e:, i:/ and the back vowels /o:, u:/, which was underpinned with a statistical analysis. Nevertheless, the discrimination between vowels is limited due to pitch dependence and at higher pitch due to a vowel migration of classical singing, which has been discussed in literature on vowel intelligibility [35,36]. This and the high variability in the data limits the applicability of simplified metrics for vowel identification in performance analysis, however the results give valuable insight on singing voice directivity. The voice directivity characteristics indicate a minor trend of increased voice directivity for the singers from breathy to pressed phonation mode. The current results show that the singing voice directivity is strongly influenced by the following components: mouth opening, singers morphology, pitch (spectral composition) and its interpretation also depends on the chosen method of analysis.

Author Contributions: Conceptualization, M.B. and M.F.; methodology, M.B. and M.F.; validation, M.B. and M.F.; formal analysis, M.B., M.F. and A.S.; investigation, M.B. and M.F.; data curation, M.B; writing—original draft preparation, M.B.; writing—review and editing, M.B.; visualization, M.B.; supervision, M.F. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of University of Music and Performing Arts, Graz.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Singers were payed an expense allowance.

Data Availability Statement: Data is available under https://phaidra.kug.ac.at/o:127031, accessed on 26 September 2022.

Acknowledgments: Special thanks to Thomas Musil for his work on the pd patch for the enhanced measurement system and to Stefan Warum for the help designing the DCMA.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

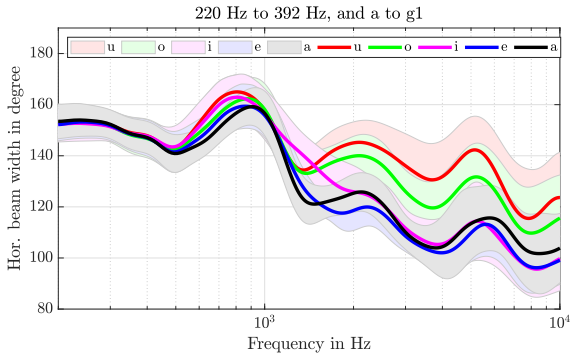
HDI	Horizontal directivity index
VDI	Vertical directivity index
IQR	Inter-quartile range
DCMA	Double circle microphone array
LTAS	Long-term averaged spectra
pd	Pure Data
r_E	Energy vector

Appendix A

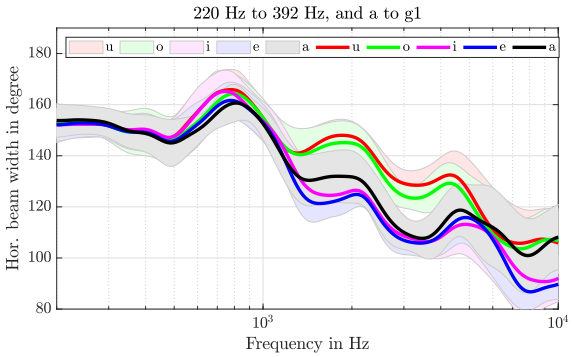
In addition to the results for the directivity metrics in Section 3.1, we present the corresponding results for the introduced beam width metric in Figure A1. Furthermore, we present the biserial correlation coefficients (effect size) for the vowel pairs for each plane for the beam width and directivity index for each pitch in Tables A1–A4.

Table A1. Biserial correlation coefficients for the horizontal and vertical beamwidths at each pitch for the female singers (samples $n = 18$). Horizontal and vertical biserial correlation coefficients are listed at the top and bottom in each cell. Bold letters indicate that there is a significant difference ($p > 0.05$) between the vowel groups listed in the first and second column.

v1	v2	a	h	c1	d1	e1	f1	g1	a1	h1	c2	d2	e2	f2	g2	a2
a	*e*	**0.9** **0.9**	**0.9** **0.8**	0.7 0.5	0.6 0.4	0.5 0.4	0.4 0.1	0 0.5	0.1 **1**	0.8 0.5	**0.9** 0.2	0.8 0.6	0.2 0.1	0.3 0	0.1 0.2	0.4 0.5
a	*i*	0.4 0.1	0.6 0.2	0.5 0	0.3 0.1	**0.9** 0.1	0.3 0.6	0.7 0.5	0.7 0	0.7 0.4	0.6 0.5	0.3 0.1	0.1 0.3	0.1 0.1	0.4 0.4	0.3 0.8
a	*o*	**1** **1**	**1** **1**	**0.9** **1**	**0.9** **1**	**0.9** **1**	**0.9** **1**	**0.9** **1**	0.8 **0.8**	0.4 **0.9**	0.6 **0.9**	0.7 **0.9**	0.4 **0.8**	0.1 0.3	0.5 0.5	0.3 0.7
a	*u*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**0.9** **1**	0.8 **1**	**0.9** **1**	0.7 **1**	0.2 0.5	0.4 0.5	0.2 0.8
e	*i*	**0.9** **1**	0.4 **0.9**	0.3 **0.9**	0.2 0.3	0.5 0.1	0.1 **0.8**	0.6 **0.9**	0.6 0.6	0.2 0.7	0 0.5	0.4 0.6	0.3 0.5	0.1 0.1	0.4 0.2	0.1 0.3
e	*o*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**0.9** **1**	**0.9** **1**	0.7 **1**	**0.8** **1**	0.8 **0.9**	**1** **1**	**0.9** **1**	0.6 **0.9**	0.3 0.4	0.3 0.4	0.3 0.2
e	*u*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	0.6 **0.9**	0.1 0.5	0.4 0.4	0.3 0.3
i	*o*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **0.9**	**1** **1**	0.7 0.7	**0.9** 0.6	0.6 0.6	0.4 0.7	0.2 0.3	0.5 0.3	0.2 0
i	*u*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**1** **1**	0.8 **1**	0.4 0.8	0 0.6	0.5 0.3	0 0.2
o	*u*	**1** **1**	**1** **0.9**	**0.9** **0.9**	**0.9** **1**	**0.9** **0.9**	**0.9** **0.9**	**0.8** **1**	0.8 **0.9**	0.6 **1**	0.4 **0.9**	0.6 **0.9**	0.1 0.1	0.1 **0.6**	0.1 0.2	0.1 0.4

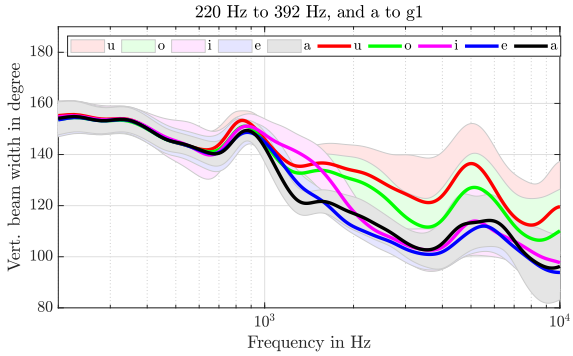


(**A**) Female singers (a to g1).

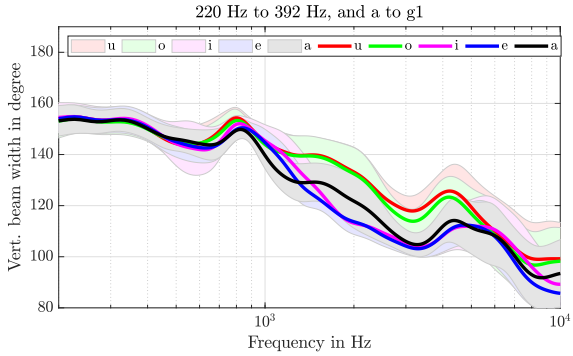


(**B**) Male singers (a to g1).

Figure A1. *Cont.*



(**C**) Female singers (a to g1).



(**D**) Male singers (a to g1).

Figure A1. Mean and standard deviation for the horizontal and vertical beam width in degree averaged over pitch for the female singers and male singers for all five vowels and all phonation modes.

Table A2. Biserial correlation coefficients for the horizontal and vertical directivity indexes at each pitch for the female singers (samples $n = 18$). Horizontal and vertical biserial correlation coefficients are listed at the top and bottom in each cell. Bold letters indicate that there is a significant difference ($p > 0.05$) between the vowel groups listed in the first and second column.

v1	v2	a	h	c1	d1	e1	f1	g1	a1	h1	c2	d2	e2	f2	g2	a2
a	*e*	0.3 0.1	0.1 0.4	0.2 0.2	0.1 0.1	0.3 0.3	0.6 0.3	0.3 0.2	0.5 0	0.3 0.7	0.1 0.6	0.1 0.6	0.6 0.2	0.7 0.6	0.2 0.1	0.6 0.1
a	*i*	0.2 0.6	0.2 0.2	0 0.2	0 0.2	0.3 0.4	0.1 0.3	0.3 0.6	0 0.2	0.1 0.3	0.1 0.5	0.5 0.4	0.7 0.2	0.6 0.3	0.3 0.1	0.6 0.1
a	*o*	**1** **0.9**	**0.9** **1**	**1** **1**	**0.9** **0.9**	0.5 0.5	**0.9** 0.8	**0.9** 0.6	**0.9** 0.7	**0.9** 0.7	**0.9** 0.4	**1** 0.5	0.7 0.3	0.4 0.4	0.5 0.3	0.4 0.1
a	*u*	**1** **1**	**1** **1**	**1** **0.9**	**1** **1**	**0.6** **0.6**	**1** **1**	**1** **1**	**1** **1**	**1** 0.5	**1** 0.6	**1** **0.9**	0.8 0.4	0.8 0.1	0.1 0	0.2 0.2
e	*i*	0.7 **0.9**	0.2 **0.8**	0.1 0.5	0.2 0.1	0.5 0.4	0.6 0.7	0.5 0.7	0.5 0.6	0.1 0.1	0.1 0.3	0.5 0.7	0.7 0.2	0.1 0.2	0.4 0.2	0.2 0.1
e	*o*	**1** **1**	**1** **0.9**	**1** **0.9**	**1** 0.8	0.3 0.4	0.5 0.7	0.4 0.7	0.5 0.7	0.8 **0.9**	**0.9** 0.8	**1** 0.8	0.1 0.3	0.7 0.7	0.6 0.4	0.3 0
e	*u*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	0.5 **0.6**	**1** **1**	**0.9** **1**	**1** **1**	**1** **1**	**1** **0.9**	**1** **0.9**	0.3 0	0.4 0.5	0.4 0.1	0.4 0.1
i	*o*	**1** **1**	**1** **0.8**	**1** **0.8**	**0.9** 0.8	**0.6** 0.2	**1** 0.5	**0.9** 0.1	**0.9** 0.1	0.8 0.6	**0.9** 0.8	0.2 0.1	0.2 0.2	0.7 0.7	0.6 0.3	0.4 0
i	*u*	**1** **1**	**1** **1**	**1** **1**	**1** **1**	**0.6** 0.5	**1** **1**	**1** 0.8	**1** **0.9**	**0.9** 0.8	**1** **0.9**	0.5 0.6	0.1 0.1	0.3 0.3	0.4 0	0.4 0.3
o	*u*	**0.9** **0.8**	**0.9** **0.9**	0.8 0.3	**0.8** 0.5	0.4 0.4	**0.9** 0.6	**0.9** **0.9**	**0.8** 0.6	0.2 0.3	0.8 0.3	0.5 0.2	0.3 0.4	0.7 0.6	0.3 0.2	0.2 0.5

Table A3. Biserial correlation coefficients for the horizontal and vertical beamwidths at each pitch for the male singers (samples n_{H-e_1} = 12, samples $n_{f_1-a_1}$ = 9). Horizontal and vertical biserial correlation coefficients are listed at the top and bottom in each cell. Bold letters indicate that there is a significant difference ($p > 0.05$) between the vowel groups listed in the first and second column.

v1	v2	H	c	d	e	f	g	a	h	c1	d1	e1	f1	g1	a1
a	*e*	**0.8**	0.7	0.8	**0.9**	**1**	**0.8**	**1**	**0.8**	0.6	**0.8**	0.7	0.4	0.5	0.3
		1	0.6	**1**	**1**	**1**	**1**	0.8	**1**	**1**	**1**	**1**	**0.6**	0.5	0.4
a	*i*	0.7	0.4	0.5	0.5	0.6	0.2	0.5	0.3	0.3	0.7	0.8	**0.6**	0.5	0.2
		0.9	0.2	0.7	0.3	**0.8**	0.4	0.6	0.7	**1**	**1**	**1**	**0.6**	0.4	0.3
a	*o*	**1**	**1**	**1**	**1**	**0.9**	**1**	**1**	**0.9**	**0.9**	**0.8**	0.6	0.1	0.2	0
		0.6	**0.9**	0.6	0.5	0.4	0.7	0.5	0.4	0.3	0.2	0.1	0.2	0.4	0.4
a	*u*	**1**	**1**	**1**	**1**	**0.9**	**1**	**1**	**0.9**	**0.9**	**0.8**	0.5	0	0	0.3
		0.8	**0.8**	0.7	**0.9**	0.7	0.7	0.4	0.4	0	0.4	0.1	0.3	0.3	0.1
e	*i*	**0.9**	0.6	0.6	0.6	0.5	0.7	**0.8**	0.5	0.4	0.5	0.2	0.2	0.1	0.2
		0.9	0.2	0.5	**0.8**	**0.8**	0.7	0.7	**0.9**	0.7	0.5	0.8	0.2	0	0.1
e	*o*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.6**	0.4	0.3
		1	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	**1**	**1**	**1**	0.5	0.4	0.1
e	*u*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.6**	0.6	0.4
		1	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	0.7	**1**	**1**	0.4	0.4	0.2
i	*o*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	**0.9**	**0.9**	**1**	**0.6**	0.6	0.2
		1	**1**	**1**	0.5	**0.8**	**0.9**	0.7	0.6	0.8	**0.8**	0.7	0.5	0.3	0.2
i	*u*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.8**	**1**	**1**	**0.6**	0.6	0.4
		1	**0.9**	**1**	**1**	**1**	**0.9**	0.7	0.5	0.4	**0.9**	**0.9**	0.3	0.5	0.1
o	*u*	0.5	**0.8**	0.7	0.7	**0.8**	0.1	**0.8**	0.5	0.2	0.3	0.3	0	0.4	0.4
		0.6	0.4	0.4	0.7	0.7	0	0	0	0	0.5	0.1	0.1	0	0.2

Table A4. Biserial correlation coefficients for the horizontal and vertical directivity indexes at each pitch for the male singers (samples n_{H-e_1} = 12, samples $n_{f_1-a_1}$ = 9). Horizontal and vertical biserial correlation coefficients are listed at the top and bottom in each cell. Bold letters indicate that there is a significant difference ($p > 0.05$) between the vowel groups listed in the first and second column.

v1	v2	H	c	d	e	f	g	a	h	c1	d1	e1	f1	g1	a1
a	*e*	**1**	**1**	**0.9**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	**0.6**	**0.6**	**0.6**
		0.9	**0.8**	0.6	**0.9**	0.7	**0.8**	0.7	**0.8**	0.7	0.6	0.8	0.3	0.4	0.1
a	*i*	**0.8**	0.3	0.5	0.4	0.6	0.5	0.6	0.8	**0.9**	**1**	**0.9**	**0.6**	**0.6**	0.6
		0.8	0.4	0.1	0.2	0.6	0.3	0.4	0.2	0.6	0.4	0.4	0.2	0.1	0.2
a	*o*	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	0.5	0.3	0.4	0.4	0.3	0.4	0.2
		1	**1**	**1**	**1**	**1**	**1**	**0.9**	**0.9**	**0.8**	0.8	0.6	0.1	0.1	0.2
a	*u*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.8**	0.5	0.2	0.2	0.3	0.4	0.1
		1	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	**0.8**	0.6	0.6	0.1	0.1	0.5
e	*i*	**0.9**	**0.9**	**0.8**	**0.8**	**0.8**	0.6	0.7	0.5	0.4	0.1	0	0.3	0.1	0
		0.8	**0.9**	0.7	**0.9**	**0.8**	**0.8**	0.5	0.6	0.3	0.6	0.7	0.3	0.6	0.4
e	*o*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.6**	0.5	0.3
		1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0.5	0.4	0.3

Table A4. *Cont.*

v1	v2	H	c	d	e	f	g	a	h	c1	d1	e1	f1	g1	a1
e	*u*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0.5	0.5	0.3
		1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0.5	0.4	0.4
i	*o*	**1**	**1**	**1**	**0.9**	**0.9**	**1**	**0.9**	**0.9**	**1**	**1**	**1**	**0.6**	0.5	0.2
		1	**1**	**1**	**1**	**1**	**1**	**0.9**	**1**	**0.9**	**0.9**	**1**	0.2	0.3	0.1
i	*u*	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.6**	0.5	0.3
		1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.9**	**0.9**	**1**	0.3	0.2	0.3
o	*u*	**0.9**	**0.9**	**0.9**	0.7	**0.8**	0.4	0.6	0.7	0.3	0.2	0.2	0.2	0.1	0.2
		1	**0.8**	**0.8**	0.6	**0.8**	0.6	0.7	0.3	0.2	0	0.3	0.4	0.1	**0.6**

References

1. Flanagan, J.L. Analog Measurements of Sound Radiation from the Mouth. *J. Acoust. Soc. Am.* **1960**, *32*, 1613–1620. [CrossRef]
2. Huopaniemi, J.; Kettunen, K.; Rahkonen, J. Measurement and modeling techniques for directional sound radiation from the mouth. In Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA'99 (Cat. No.99TH8452), New Paltz, NY, USA, 20–20 October 1999; pp. 183–186. [CrossRef]
3. Halkosaari, T.; Vaalgamaa, M.; Karjalainen, M. Directivity of Artificial and Human Speech. *J. Audio Eng. Soc.* **2005**, *53*, 620–631.
4. Chu, W.T.; Warnock, A.C. *Detailed Directivity of Sound Fields Around Human Talkers*; Technical report-NRC-CNRC Publications Archive; National Research Council Canada: Ottawa, ON, Canada, 2002. [CrossRef]
5. Monson, B.B.; Hunter, E.J.; Story, B.H. Horizontal directivity of low- and high-frequency energy in speech and singing. *J. Acoust. Soc. Am.* **2012**, *132*, 433–441. [CrossRef]
6. Kocon, P.; Monson, B.B. Horizontal directivity patterns differ between vowels extracted from running speech. *J. Acoust. Soc. Am.* **2018**, *144*, EL7–EL12. [CrossRef]
7. Pörschmann, C.; Arend, J.M. Investigating phoneme-dependencies of spherical voice directivity patterns. *J. Acoust. Soc. Am.* **2021**, *149*, 4553. [CrossRef]
8. Marshall, A.H.; Meyer, J. The directivity and auditory impressions of singers. *Acta Acust. United Acust.* **1985**, *58*, 130–140.
9. Kob, M.; Jers, H. Directivity measurement of a singer. *J. Acoust. Soc. Am.* **1999**, *105*, 1003.
10. Cabrera, D.; Davis, P.J. Vocal directivity measurements of eight opera singers. In *18th International Congress on Acoustics*; Kyoto, Japan, 4–9 April 2004; pp. 505–506.
11. Katz, B.; D'Alessandro, C. Directivity Measurements of the Singing Voice. In Proceedings of the 19th International Congress on Acoustics, Madrid, Spain, 2–7 September 2007; pp. 45–50.
12. Cabrera, D.; Davis, P.J.; Connolly, A. Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment. *J. Voice* **2011**, *25*, e291–e303. [CrossRef]
13. Boren, B.B.; Roginska, A. Sound radiation of trained vocalizers. *Proc. Meet. Acoust.* **2013**, *19*, 035025. [CrossRef]
14. Blandin, R.; Brandner, M. Influence of the vocal tract on voice directivity. In Proceedings of the 2019 23rd International Congress on Acoustics—ICA, Aachen, Germany, 9–13 September 2019; Deutsche Gesellschaft für Akustik e.V.: Berlin, Germany, 2019; pp. 1795–1801.
15. Brandner, M.; Blandin, R.; Frank, M.; Sontacchi, A. A pilot study on the influence of mouth configuration and torso on singing voice directivity. *J. Acoust. Soc. Am.* **2020**, *148*, 1169–1180. [CrossRef]
16. Molloy, C.T. Calculation of the Directivity Index for Various Types of Radiators. *J. Acoust. Soc. Am.* **1948**, *20*, 387–405. [CrossRef]
17. Tylka, J.G.; Sridhar, R.; Choueiri, E. A database of loudspeaker polar radiation measurements. In Proceedings of the Audio Engineering Society Convention 139, New York, NY, USA, 29 October–1 November 2015.
18. Gerzon, M.A. General metatheory of auditory localisation. In *Audio Engineering Society Convention*; Audio Engineering Society: New York, NY, USA, 1992.
19. Kadiri, S.R.; Alku, P.; Yegnanarayana, B. Analysis and classification of phonation types in speech and singing voice. *Speech Commun.* **2020**, *118*, 33–47. [CrossRef]
20. Zhang, Z. Mechanics of human voice production and control. *J. Acoust. Soc. Am.* **2016**, *140*, 2614–2635. [CrossRef]
21. Brandner, M.; Frank, M.; Rudrich, D. DirPat—Database and viewer of 2D/3D directivity patterns of sound sources and receivers. In Proceedings of the Audio Engineering Society Convention 144, Milan, Italy, 23–26 May 2018.
22. Frank, M.; Rudrich, D.; Brandner, M. Augmented practice-room—Augmented acoustics in music education. In *Fortschritte der Akustik, DAGA; Deutsche Gesellschaft für Akustik*: Berlin, Germany, 2020; pp. 151–154.
23. Klanjscek, N.; David, L.; Frank, M. Evaluation of an e-learning tool for augmented acoustics in music education. *Music Sci.* **2021**, *4*, 20592043211037511. [CrossRef]
24. Meyer-Kahlen, N.; Rudrich, D.; Brandner, M.; Wirler, S.; Windtner, S.; Frank, M. Diy modifications for acoustically transparent headphones. In *Audio Engineering Society Convention 148*; Audio Engineering Society: New York, NY, USA, 2020.

3 Perception of Voice Directivity

Several studies in literature investigated the voice directivity patterns of human speech, but only a few investigated the perception of voice directivity and the magnitude of change necessary to be perceived by a listener. In order to quantify how well changes in voice directivity are perceivable, two works are presented. One examines the perception of voice directivity at a listening position somewhere in a room (see Section 3.1) and the second investigates the case where the listener position coincides with the speaker or singer's position, meaning when listening to one's own voice (see Section 3.2). The findings demonstrate that a listener in the room perceives the most pronounced difference at the critical distance when oriented on-axis toward the sound source. Changes in voice directivity when listening to someone's own voice are hardly perceivable due to the high signal to noise ratio present at the ears and due to the close distance to the mouth.

3.1 Perceptual Evaluation of Spatial Resolution in Directivity Patterns

This work was published as:

M. Frank and **M. Brandner**. Perceptual Evaluation of Spatial Resolution in Directivity Patterns. *Proceedings of the DAGA*, 46:74–77, Vienna, 2019.

The idea and concept of this article were outlined by the first author and me. The first author wrote the original draft of the manuscript with periodical contributions from me. The revision and editing was done by the first author and me. I did help to prepare and conduct the listening test.

DAGA 2019 Rostock

Perceptual Evaluation of Spatial Resolution in Directivity Patterns

Matthias Frank[1] and Manuel Brandner[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria, Email: frank@iem.at*

Introduction

Plausible and authentic auralization of sound sources in rooms benefits from the incorporation of source and receiver directivity to control the direct-to-reverberant energy ratio, and thus the perceived distance [1, 2]. While receiver directivity is typically well represented by measured head-related transfer functions, source directivity requires more measurement effort, especially for musical instruments. However, there exist surround microphone arrays of 64 microphones to capture source directivity in a resolution up to 7$^{\text{th}}$ order spherical harmonics [3]. Even if the directivity pattern is not of high spatial resolution, high orders are sometimes necessary to compensate for imprecise centering [4, 5].

Clearly, the computational effort for auralization with high orders is high and can make real-time processing very challenging, in particular when using multiple sources simultaneously. Moreover, typical compact spherical loudspeaker arrays to play back directional sources are limited to 3$^{\text{rd}}$ order [6]. As a consequence, it is desirable to reduce the spatial resolution of directivity patterns. Still, little is known about the perceptual impact of such reduction and what the best strategy for reduction is, i.e. which parameters of the directivity pattern are most important to preserve.

This contribution tries to take a step forward in answering these questions. As a simplified model, we assume that every arbitrary directivity pattern can be composed of multiple directional components and a diffuse component. The directional components are characterized by parameters, such as direction, beam width, and side lobe suppression. The resulting phase of the combined directional components is assumed to be perceptually irrelevant when playing the source in a reverberant environment. The diffuse component distributes decorrelated sound equally into all directions and its level is adjusted to complement the directional components in a way that the overall sound power is preserved in comparison to the high order directivity pattern.

In a listening experiment, we investigate the perceptual impact of order reduction for both a generic directional beam and a diffuse source. The experiment uses speech and noise as stimuli and is performed at two listening positions in a simulated room. The experimental results are finally related to technical measures to investigate which parameters of the source directivity are most important to preserve. This helps to develop an efficient reduction strategy.

Setup and Conditions

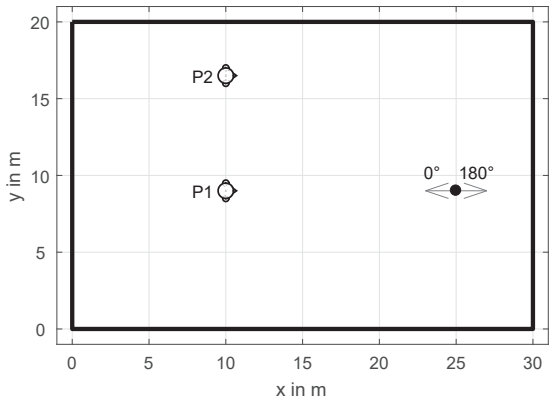


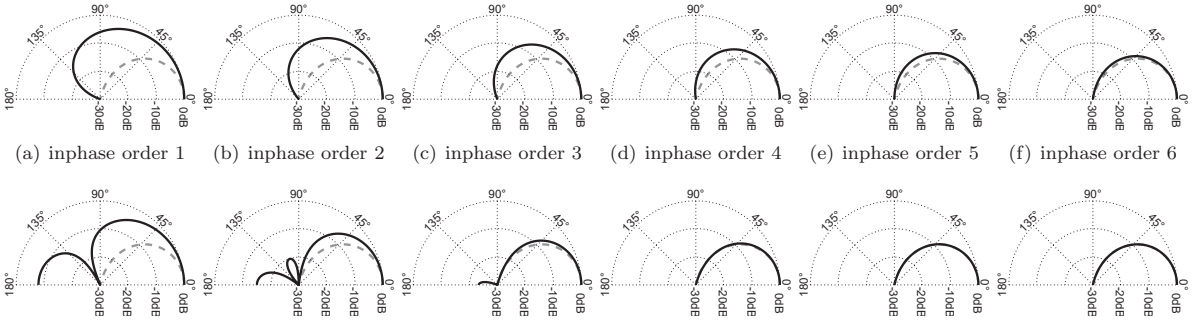
Figure 1: Position of listener and source in the horizontal cross section of the simulated room.

The simulated room had a size of 30 m × 20 m × 10 m, cf. Figure 1. The reverberation time for frequencies between 200 Hz and 2 kHz was 1.9 s and it doubled/halved for frequencies below 100 Hz and above 4 kHz, respectively. That resulted in a critical distance of 3.2 m for mid frequencies. The simulation employed a 7$^{\text{th}}$-order image-source model (236 reflections) from the IEM `RoomEncoder` VST plug-in[1]. Headphone playback employed 7$^{\text{th}}$-order head-tracked [7] binaural Ambisonics [8] using the IEM `BinauralDecoder`.

The source was positioned 4 m above the floor and the two listening positions were at a height of 2 m to recreate typical concert conditions. Listening position P1 was positioned exactly in the frontal direction of the source to provoke changes in the direct-to-reverberant energy ratio mainly, whereas P2 had an angle of 30° to the source to study the influence of an increased beam width at low orders.

The reference for the directional beam was a 7$^{\text{th}}$-order inphase [9] design facing exactly at P1 (0°) and away from it (180°), respectively. The limitation to 7$^{\text{th}}$ order was done due to practical reasons, as the largest available microphone array to measure simultaneously has 64 microphones and the VST plug-in is also limited to 64 channels. The second orientation resulted in no direct sound from the source at P1, as the inphase design has a null at the back, cf. gray dashed directivity pattern in Figure 2. The distance between the source and P1 was chosen to be the effective critical distance of the reference beam. Note that the orientation angle of the source in the plug-in is reversed in comparison to this contribution.

[1] freely available at `plugins.iem.at`



(a) inphase order 1 (b) inphase order 2 (c) inphase order 3 (d) inphase order 4 (e) inphase order 5 (f) inphase order 6

(g) truncated order 1 (h) truncated order 2 (i) tunrcated order 3 (j) truncated order 4 (k) truncated order 5 (l) truncated order 6

Figure 2: Directivity patterns of beams in the experiment; gray dashed line indicates 7^{th}-order inphase beam as reference.

Two different strategies were employed to reduce the order of the directional beam: (I) appropriate inphase design for the reduced order, resulting in beams with strongly increasing width of the main lobes while preserving nulls at the back and (II) simple truncation that better preserves the width of the main lobe at the cost of increasingly strong side lobes at the back of the pattern, cf. Figure 2.

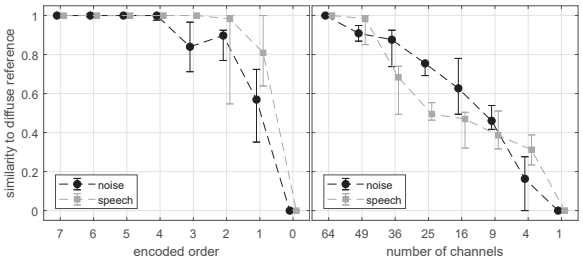
The diffuse source employed a 64×64 feedback delay network (FDN) [10, 11] to generate multiple decorrelated signals out of a single input using the IEM `FDNReverb` plug-in. The plug-in was set to a reverberation time of 1 s (0.3 s above 2 kHz) and a fade-in time of 0.1 s to ensure maximally decorrelated output signals while keeping the reverberation by the FDN clearly below the reverberation time of the room. There were two strategies to spatially distribute the 64 output signals on the diffuse source: (A) encoding the 64 signals at 64 position equally distributed on a sphere and varying the order from 0 to 7 (reference) and (B) selecting a set of output channels ranging from 1, 4, 9 to 64 (reference) and apply them as spherical harmonic channels directly.

The experiment employed two different sounds: (α) continuous pink noise for maximum sensitivity to coloration and loudness and (β) male English speech [12] that facilitates better spatial perception and familiarity.

Overall, there were 20 = 2 (sounds) × 2 (reduction strategy for directional beam) × 2 (orientation of the directional beam) × 2 (listening positions for directional beam) + 2 (diffuse strategy) trials with 8 (0^{th} to 7^{th} order or 1, 4, 9 to 64 channels) stimuli each in multi-stimulus comparison. The listeners task was to compare the similarity of the 8 stimuli to the corresponding reference on a continuous scale from *very different* to *identical*. Each of the 10 experienced listeners (average age 32 years) spent about 33 min on the entire experiment.

Results

The results for the diffuse source revealed that independent of the strategy, the similarity to the reference increases with the encoding order and number of FDN output channels, cf. Figure 3. The encoding strategy requires orders of 4 and 1 to generate perceptually indistinguishable results (Wilcoxon signed rank test with Bonferroni-Holm correction) to the reference for noise and speech, respectively. Reducing the number of FDN output channels has a stronger perceptual effect, as this strategy would require 64 and 49 channels (equivalent to orders of 7 and 6).

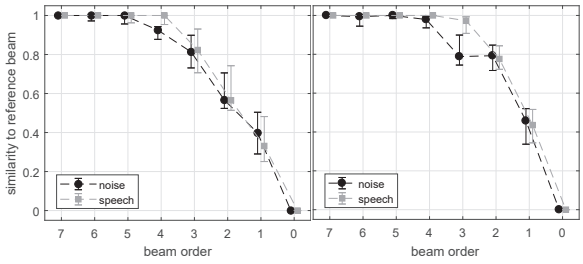


(a) 64 encoded FDN channels (b) number of FDN channels

Figure 3: Medians and 95% confidence intervals of perceived similarity for diffuse source.

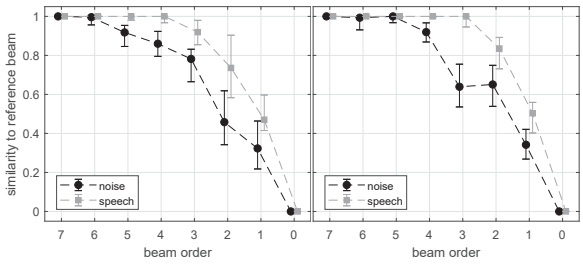
The results indicate that the number of FDN output channels is crucial for the diffuseness of the source while the spatial resolution for distributing the 64 channels on the surface of the source is not important for decorrelation as long as the spherical harmonics order exceeds 0. However, as the results for noise show, coloration-less reproduction may require higher orders.

For the directional beams, similarity increases with the order, cf. Figures 4 and 5. For each sound, reduction strategy, and source orientation, the minimum required order to be indistinguishable from the reference is shown in Table 1. As for the diffuse source, higher resolution is required for the reproduction of noise. The truncation strategy seems to be more efficient, especially for noise at listening position P2, where the beam is not facing directly at or away from the listener. This finding gives a hint that preserving the width of the main lobes is more important than the exact reproduction of the nulls.
In order to derivate rough thresholds for perceptual differences, i.e. similar to just noticeable differences (JNDs), the next section calculates some technical measures and relates them to the experimental results for the directional beams.



(a) inphase beams (b) truncated beams

Figure 4: Medians and 95% confidence intervals of perceived similarity summarizing 0° and 180° beam directions at listening position P1.



(a) inphase beams (b) truncated beams

Figure 5: Medians and 95% confidence intervals of perceived similarity summarizing 0° and 180° beam directions at listening position P2.

Table 1: Minimum required order to be indistinguishable from reference at 5% level with Bonferroni-Holm correction.

	list. pos. P1		list. pos. P2	
beam	noise	speech	noise	speech
inphase 0°	5	4	7	4
inphase 180°	5	3	6	3
truncated 0°	4	4	5	3
truncated 180°	5	3	5	3

Technical Measures

The first kind of technical measures is independent of the room and solely depends on the beam itself. The measures are:

- side lobe: level of the strongest side lobe in dB,
- width: aperture angle of the cap exceeding -6 dB relative to the maximum in °,
- F/B-R_{40}: front-to-back ratio in dB, with lower dynamic limitation at -40 dB relative to the maximum,
- F/B-R_{25}: front-to-back ratio in dB, with lower dynamic limitation at -25 dB relative to the maximum.

Table 2 presents the values for the reference beam and the two reduction strategies at different orders. Values for orders that resulted in indistinguishable experiment results for speech and noise are printed in italics and bold, respectively.
The minimum required order for the truncation strategy when facing the listener at P1 was found to be 5 and 3, respectively, cf. Table 1. Thus for noise and speech, a side lobe attenuation of 49.1 dB and 23.4 dB is enough. This finding indicates that a more precise preservation of nulls in the beam pattern are not necessary. The F/B-R_{40} reflects this limitation of the side lobe attenuation and reveals that a difference of 1.1 dB is perceptually irrelevant for noise. The difference when further reducing the order to 4 (inphase) and 3 (truncation) results in a clear increase of at least 4.4 dB. Similarly for speech, differences in F/B-R_{25} of 0.8 dB are tolerable, while the next lower order results in a difference of at least 2.5 dB. Interestingly, for both noise and speech, there is a clear point of discontinuity in the F/B-R_{40} and F/B-R_{25} differences at the minimum required order.
The differences in beam width are visible for the 0° direction at P2. For noise, a difference of 3° was not tolerated, whereas for speech 23° or 32% were tolerated.

Table 2: Side lobes, beam width, and front-to-back energy ratios of the tested beams. Values that resulted in indistinguishable results for speech and noise are printed in italics and bold, respectively.

beam	side lobe in dB	width in °	F/B-R_{40} in dB	F/B-R_{25} in dB
inphase 7	$-\infty$	71	34.8	19.8
inphase 6	$-\infty$	*77*	**35.0**	*20.1*
inphase 5	$-\infty$	*84*	**34.0**	*20.5*
inphase 4	$-\infty$	*94*	30.4	*20.9*
inphase 3	$-\infty$	108	24.7	*20.6*
inphase 2	$-\infty$	131	18.1	17.3
inphase 1	$-\infty$	180	10.8	10.8
0	0	360	0	0
truncated 6	**−70.7**	**71**	**34.8**	*19.8*
truncated 5	**−49.1**	**72**	**34.8**	*19.8*
truncated 4	*−34.2*	*74*	**33.7**	*19.9*
truncated 3	*−23.4*	*81*	24.9	*20.1*
truncated 2	-15.1	99	16.1	15.9
truncated 1	-8.0	147	9.5	9.5

View publication stats

Table 3: Direct-to-reverberant energy ratio of the tested beams in dB for both listening positions. Values that resulted in indistinguishable results for speech are printed in italics.

	list. pos. P1		list. pos. P2	
beam	180°	0°	180°	0°
inphase 7	$-\infty$	-0.2	$-\infty$	-7.3
inphase 6	$-\infty$	*−0.7*	$-\infty$	*−7.5*
inphase 5	$-\infty$	*−1.3*	$-\infty$	*−7.9*
inphase 4	$-\infty$	*−2.0*	$-\infty$	*−8.4*
inphase 3	$-\infty$	-2.8	*−80.3*	-9.0
inphase 2	$-\infty$	-3.9	-56.2	-9.8
inphase 1	-50.1	-5.3	-32.8	-10.8
0	-9.1	-9.1	-13.8	-13.8
truncated 6	*−71.6*	*−0.2*	*−81.3*	*−7.3*
truncated 5	*−49.4*	*−0.2*	*−62.4*	*−7.3*
truncated 4	*−34.1*	*−0.4*	*−59.0*	*−7.2*
truncated 3	*−23.4*	-0.9	*−40.6*	*−7.5*
truncated 2	-16.1	-2.3	-25.5	-8.5
truncated 1	-10.7	-4.5	-16.5	-10.1

The second kind of technical measure is the direct-to-reverberant energy ratio and it depends on the combination of the beam pattern, its orientation, and the room. For the 180° orientation at P1, the resulting values are obviously similar to the level of the strongest side lobe in Table 2. The same tendency can be found at P1. The results indicate that the direct-to-reverberant energy ratio has a lower perceptual limit around -23 dB for speech. For reference values around 0 dB, i.e. for the 0° orientation at P1, differences exceeding 1.8 dB are perceivable. This difference agrees with the JNDs found in literature [13]. For the same orientation at P2, where the beam is not facing the listener, sensitivity increases and so the difference to the reference must not exceed 1.1 dB.

Conclusion

This contribution investigated the perceptual effect of reducing the spatial resolution of 7^{th}-order diffuse and directional sources in a virtual environment. In general, listeners were more sensitive when listening to pink noise in comparison to speech. For the diffuse source, a high number of decorrelated FDN outputs was crucial, whereas a 1^{st}-order spatial resolution was sufficient for speech.
The minimum required order of directional beams to be indistinguishable from the 7^{th}-order inphase reference was around 3 for speech and 5 for noise. The reduction of the order by simple truncation yielded better results than the inphase design of the same order, indicating that the exact preservation of nulls is perceptually less relevant than the approximation of the beam width. Differences in beam width of 23° or 32% were tolerable for speech, while differences of 3° or 4% were perceivable for noise. Differences in the front-to-back ratio were imperceivable below about 1 dB, as for the direct-to-reverberant ratio for a reference around 0 dB. For a reference at $-\infty$ dB, a beam with a ratio of -23 dB was not perceived differently.

References

[1] A. Kolarik, S. Cirstea, and S. Pardhan, "Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3395–3398, 2013.

[2] F. Wendt, F. Zotter, M. Frank, and R. Höldrich, "Auditory distance control using a variable-directivity loudspeaker," *MDPI Applied Science*, vol. 7, no. 7, 2017.

[3] F. Hohl, "Kugelmikrofonarray zur Abstrahlungsvermessung von Musikinstrumenten," Master's thesis, TU Graz, 2009.

[4] D. Deboy, "Tangential intensity algorithm for acoustic centering," in *Fortschritte der Akustik, DAGA*, Düsseldorf, 2011.

[5] I. B. Hagai, M. Pollow, M. Vorländer, and B. Rafaely, "Acoustic centering of sources measured by surrounding spherical microphone arrays," *Journal of the Acoustical Society of America (JASA)*, vol. 130, no. 4, 2011.

[6] F. Zotter, M. Zaunschirm, M. Frank, and M. Kronlachner, "A beamformer to play with wall reflections: The icosahedral loudspeaker," *Computer Music Journal*, vol. 41, no. 3, 2017.

[7] M. Romanov, P. Berghold, M. Frank, D. Rudrich, M. Zaunschirm, and F. Zotter, "Implementation and evaluation of a low-cost headtracker for binaural synthesis," in *Audio Engineering Society Convention 142*, May 2017.

[8] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Fortschritte der Akustik - DAGA*, Munich, March 2018.

[9] J. Daniel, "Représentation des champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2001.

[10] J. Stautner and M. Puckette, "Designing multichannel reverberators," *Computer Music Journal*, vol. 6, no. 1, pp. 52–65, 1982.

[11] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *90th AES Conv., prepr. 3030*, Paris, February 1991.

[12] EBU. (2008) EBU SQAM CD: Sound Quality Assessment Material recordings for subjective tests. [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[13] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008.

3.2 Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions

This work was published as:

M. Frank and **M. Brandner**. Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions. *Proceedings of ICSA*, 5:131–135, Ilmenau. 2019. doi:10.22032/dbt.39966.

The idea and concept of this article were outlined by the first author and me. The first author wrote the original draft of the manuscript with periodical contributions from me. The revision and editing was done by the first author and me. The preparation and conducting of the listening test were done by the first author and me.

Proceedings of ICSA 2019

5th International Conference on Spatial Audio
September 26th to 28th, 2019, Ilmenau, Germany

DOI: 10.22032/dbt.39966

DOI (proceedings): 10.22032/dbt.39936

vdt Verband Deutscher Tonmeister e.V.

Full Reviewed Paper at ICSA 2019

Presented * by VDT.

Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions

Matthias Frank[1], Manuel Brandner[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria*
Email: frank@iem.at, brandner@iem.at

Abstract

The incorporation of source directivity is important for a plausible and authentic auralization. While high-resolution measurement setups and data exist, it is yet not clear how detailed the directivity information has to be measured and reproduced with regard to perception. In particular, when source and listener are at the same location, resulting in a high direct-to-reverberant energy ratio, the precise shape of the directivity pattern might not yield perceptual differences. The paper approaches this question by a listening experiment in a virtual environment with generic directivity patterns and coincident position of listener and source. The experiment compares different spatial resolutions (spherical harmonic orders) of the directivity patterns for multiple virtual listener/source positions/orientations and levels of direct sound for speech and noise. The virtual environment employs a higher-order image-source model and binaural, dynamic Ambisonic playback. The results show that the exact shape of the directivity pattern is often perceptually irrelevant, while the preservation of the direct-to-reverberant energy ratio is more important.

1. Introduction

Plausible and authentic auralization of sound sources in rooms benefits from the incorporation of source directivity and variable source orientation [1]. This is mainly due to the natural perception of distance that is controlled by the direct-to-reverberant energy ratio (DRR) [2, 3]. High-resolution measurement of source directivity is typically done with surrounding microphone arrays of up to 64 microphones at the same time [4] and directivity patterns are often represented in spherical harmonics to facilitate simple rotation. A high resolution is sometimes necessary to compensate for imprecise centering [5, 6], even for sources with low spatial resolution in their directivity patterns. Our previous study [7] revealed that perception of spatial resolution in directivity patterns is limited to spherical harmonic orders around 4 for large distances between source and receiver in a stimulated concert hall. In such cases, the DRR is typically negative.

However, for the auralization of one's own voice or when playing an instrument oneself [8–10], direct sound dominates. This paper investigates how precise directivity patterns are perceived in such cases, i.e. to which order a higher-order directivity pattern can be reduced to still be perceptually indistinguishable from a reference. The reference directivity pattern is highly directive as it appears for large brass instruments at high frequencies. The investigation employs a high level of direct sound as it appears in human speaking/singing and further, reduced levels to represent instruments with less direct sound at the player's ears. The virtual room in which the directional source is playing is simulated by an image-source model without late diffuse reverberation. These settings are chosen to simulate the most sensitive case, whereas a practical application might be less critical.

The paper first introduces setup and conditions of the listening experiment. The following section presents the experimental results. The results are subsequently compared to technical measures that are related to room acoustics and properties of the directivity patterns. Finally, the investigation is summarized and compared to our previous results in [7] for non-coincident listener and source.

© Verband Deutscher Tonmeister e.V., 2019

Proceedings of ICSA 2019

5th International Conference on Spatial Audio
September 26th to 28th, 2019, Ilmenau, Germany

2. Setup and Conditions

The parameters of the room simulation were identical to those used in [7]: The room had a size of 30 m × 20 m × 10 m and a reverberation time of 1.9 s between 200 Hz and 2 kHz, and was doubled/halved for frequencies below 100 Hz and above 4 kHz, respectively. The simulation employed a 7^{th}-order image-source model (236 reflections) implemented in the IEM `RoomEncoder` VST plug-in[1]. The headphone playback employed 7^{th}-order head-tracked [11] binaural Ambisonics [12] using the IEM `BinauralDecoder`. Note that the rotation of the source was linked to the head rotation.

The direct sound was not generated by the `RoomEncoder` plug-in, as this would result in an infinitely high level for coincident source and receiver position. Therefore, it was realized as omnidirectional sound inside the listener's head with a specific level that should correspond to direct sound level at a speaker's own ears. The level is based on a measurement of a B&K HATS 4128 using its mouth simulator, its ears, and two omnidirectional microphones at 1 m and 25 mm distance (mouth reference point, MRP) from the mouth in an anechoic chamber, respectively. Fig. 2 shows that the level at the ears is roughly 20 dB less than at the MRP. These results are similar to findings in [8] and the deviations can be explained by different distances of the MRP. The level in 1 m distance is again about 10 dB less than at the ears. A broad-band level difference of 10 dB was used to calibrate the direct sound and the image-source model for the experiment and is denoted as 0 dB direct sound level in the remainder of this paper. In order to represent instruments with less direct sound at the player's ears, reduced levels {-10, -20} dB were also evaluated.

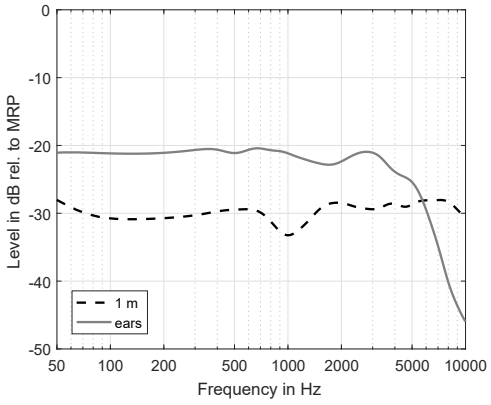


Fig. 2: Sound pressure levels in 1 m in front of the HATS and at its ears relative to the mouth reference point (MRP).

[1] freely available at plugins.iem.at

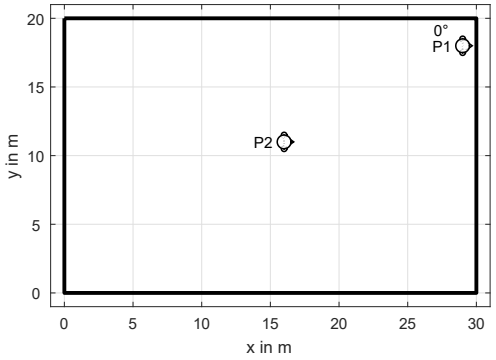
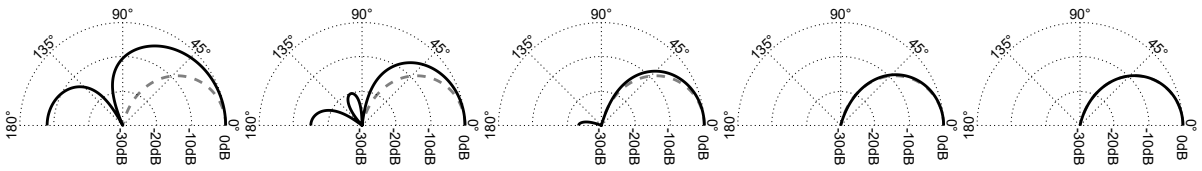


Fig. 3: Listener/source position in the horizontal cross section of the simulated room. Indicated listener/source orientation is defined as 0°.

The source and the listener were positioned coincidentally with a height of 4 m above the floor at positions P1 and P2. P1 was close to a wall to provoke a strong first reflection that could interfere with the direct sound, when the source/listener was facing the wall (0° orientation), cf. Fig. 3. In contrast, the second orientation (180°) at P1 yielded weaker reflections. For P2, which was close to the center of the room, the reflection pattern was less orientation-dependent. Thus, there was only one orientation evaluated at P2.

The reference directivity was a 7^{th}-order inphase [13] design, resulting in no side lobes and a relatively narrow main lobe, cf. Fig. 1. This directivity is similar to that of larger brass instruments, e.g. trombones or tubas, at high frequencies [14]. Typical directivity patterns of other instruments can be assumed to be less directive. In the experiment, the reference directivity pattern was reduced to orders 0 to 5 by simple truncation, as our previous study [7] revealed truncation to be perceptually better than preservation of nulls. Orders higher than 5 were excluded, as they were perceived as identical to the reference in preliminary tests. All resulting directivity patterns were diffuse-field equalized. The experiment employed two different sounds: (a) continuous pink noise for maximum sensitivity to coloration and (b) male English speech [15] that facilitates better spatial perception and familiarity.

Overall, there were 18 = 2 (sounds) × 3 ({0, -10, -20} dB direct sound level) × 3 (2 orientations at P1 + 1 orientation at P2) trials with multi-stimulus comparisons. The listeners task was to compare the similarity of the 6 (0^{th} to 5^{th} order truncation) stimuli to the corresponding 7^{th}-order reference on a continuous scale from *very different* to *identical*. Note that the playback level in each trial was adjusted reversely to the level of the direct sound in order to achieve similar loudness between the trials.



(a) truncated to order 1 (b) truncated to order 2 (c) tunrcated to order 3 (d) truncated to order 4 (e) truncated to order 5

Fig. 1: On-axis equalized directivity patterns of beams in the experiment; gray dashed line indicates 7^{th}-order inphase beam as reference.

© Verband Deutscher Tonmeister e.V., 2019

Proceedings of ICSA 2019

5th International Conference on Spatial Audio
September 26th to 28th, 2019, Ilmenau, Germany

3. Results

There were 10 experienced listeners (average age 31 years) who spent about 21 min each on the entire experiment. Based on the 10 values for each condition, the results of the experiment are presented as median values and corresponding confidence intervals in Figs. 4 and 5 for noise and speech, respectively. The gray level of the markers and lines in the figures indicates the level of the direct sound. Obviously, the similarity to the reference increases with the truncation order and also with the level of the direct sound for both sounds and all positions/orientations.

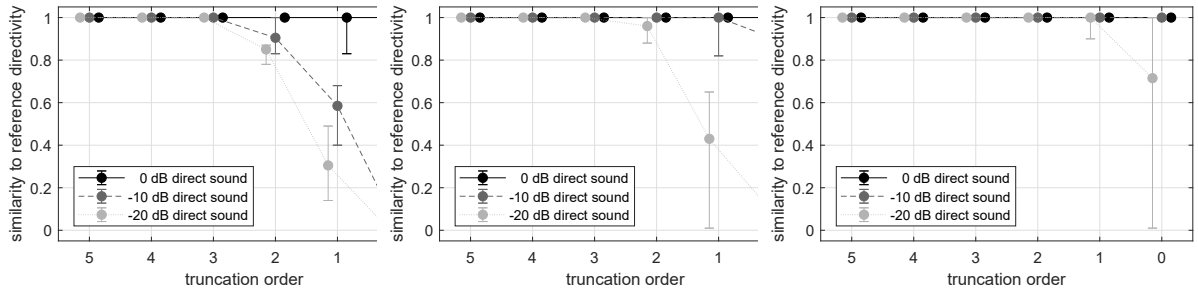
As we were interested in the spatial resolution required for perceptually indistinguishable auralization in comparison to the reference, Tab. 1 provides a suitable and easy-to-read representation of the results: For each condition, the table shows the minimum required order to yield indistinguishable results in terms of a Wilcoxon signed rank test with Bonferroni-Holm correction.

The influence of the direct sound can be seen clearly: While at the lowest level (-20 dB) orders around 2 are required, results are perceptually indistinguishable from the reference already for an order of 0 at the highest level (0 dB) for all conditions except speech at P1 and 0° orientation. This indicates that for dominant direct sound, the exact control of the reflections by the directivity pattern is not important as long as the direct-to-reverberant energy ratio is preserved. This seems to be already assured by the diffuse-field equalization of the truncated directivity patterns.

The sensitivity of the noise conditions increases with the proximity and orientation towards the walls: The central position P2 is most distant to all walls and it requires only an order of 1 or 0 for -20 dB or -10 dB direct sound, respectively. When facing the close wall at P1 and 0° orientation, orders of 3 and 2 are required for the same level of direct sound. In contrast, there is no dependency on the listener/source position and the orientation for speech, except for the increased sensitivity at P1 with 0° orientation. The increased sensitivity of noise in comparison to speech at P1 for 0° orientation and -20 dB direct sound is due to a strong comb filter. As listeners reported after the experiment, the truncation led to different strength of comb filters for noise, while it led to different level and density of reverberation for speech.

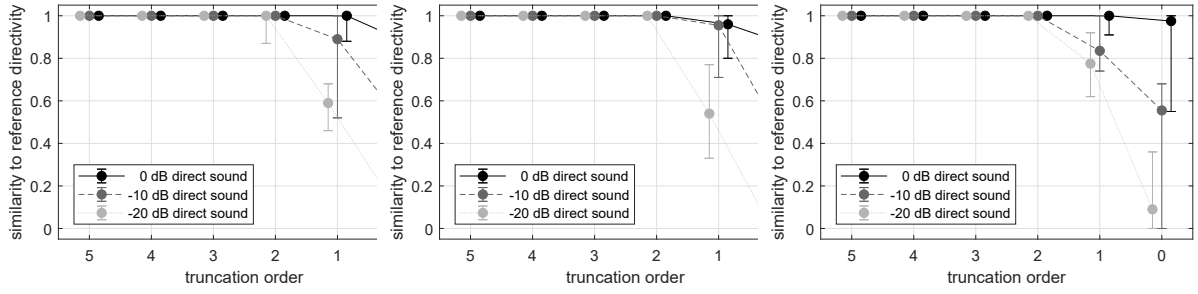
Tab. 1: Minimum required order to be indistinguishable from reference at 5% level with Bonferroni-Holm correction.

	0 dB direct sound			-10 dB direct sound			-20 dB direct sound		
sound	P1, 0°	P1, 180°	P2, 0°	P1, 0°	P1, 180°	P2, 0°	P1, 0°	P1, 180°	P2, 0°
noise	0	0	0	2	1	0	3	2	1
speech	1	0	0	1	1	1	2	2	2



(a) position P1, orientation 0° (towards the wall) (b) position P1, orientation 180° (into the room) (c) position P2, orientation 0° (into the room)

Fig. 4: Medians and 95% confidence intervals of perceived similarity to auralization using $7^{\text{th}}-$order reference directivity for noise at different listening/source positions/ and orientations for different levels of direct sound.



(a) position P1, orientation 0° (towards the wall) (b) position P1, orientation 180° (into the room) (c) position P2, orientation 0° (into the room)

Fig. 5: Medians and 95% confidence intervals of perceived similarity to auralization using $7^{\text{th}}-$order reference directivity for speech at different listening/source positions and orientations for different levels of direct sound.

© Verband Deutscher Tonmeister e.V., 2019

Proceedings of ICSA 2019

5th International Conference on Spatial Audio
September 26th to 28th, 2019, Ilmenau, Germany

4. Technical Measures

This section calculates some technical measures in order to generalize the experimental results for application on different room settings and directivity patterns.

The first kind of technical measure is the direct-to-reverberant energy ratio (DRR) and it depends on the combination of the directivity pattern, its orientation, the listener/source position, the direct sound level and the room. Note that in our calculation of DRR, the first reflections also contributed to the reverberant energy. Tab. 2 shows the resulting values in dependence of the direct sound level. Naturally, the DRR increases with the level of direct sound. The $0°$ orientation at P1 results in values about 16 dB lower than for the $180°$ orientation and P2 because it yields a strong first reflection from the nearby wall. In this case, the DRR increases for truncated orders due to a reduction of the reflection from the wall, i.e. the diffuse-field equalized directivity patterns radiate more energy into all other directions away from the wall. A similar, however weaker behavior can be seen at P2. In contrast, order truncation of the directivity pattern reduces the DRR for the $180°$ orientation at P1. Here, the lower-order patterns lead to an increase of the energy from the nearby wall that in turn reduces the DRR values.

Tab. 2 relates to the experimental results by printing values in bold that resulted in indistinguishable results for speech. For reference DRR values around 40 dB, deviations of around 4 dB were not perceivable. For values around 30 dB, deviations must not exceed 2 dB to remain perceptually irrelevant. Similar sensitivity can be found for DRR values around 0 dB. The tendency that sensitivity decreases towards higher DRR agrees with literature [16]. However, there are exceptions, where the threshold is smaller (P2, $0°$ with -20 dB direct sound: below 1 dB). This might be due to the different strategies for creating the stimuli: In [16], the direct sound was attenuated/boosted and the rest of the impulse response was kept identical. In our experiment, the modification of the directivity patterns modified the impulse response but the direct sound remained the same. In this way, we did not directly modify the level ratio between direct sound and reverberation, but the level of each reflection in the impulse response.

Tab. 3: Side lobes, beam width, and front-to-back energy ratio of the tested directivity patterns.

directivity order	side lobe in dB	width in °	F/B-R_{25} in dB
7 (ref)	$-\infty$	71	19.8
5	-49.1	72	19.8
4	-34.2	74	19.9
3	-23.4	81	20.1
2	-15.1	99	15.9
1	-8.0	147	9.5
0	0	360	0

The second kind of technical measures is independent of the room and the listener/source position because it solely depends on the directivity pattern itself. The measures are (a) side lobe: level of the strongest side lobe in dB, (b) width: aperture angle of the cap exceeding -6 dB relative to the maximum at the $0°$ direction in °, and (c) F/B-R_{25}: front-to-back ratio in dB, with lower dynamic limitation at -25 dB relative to the maximum [7].

Tab. 3 shows the above-mentioned measures for the reference directivity and the directivities truncated at different orders. For -20 dB direct sound, the minimum required order for speech was 2. In this case, a side lobe attenuation of around 15 dB was not distinguished from the reference, a widening of the beam of $28°$ or 39%, and a F/B-R_{25} difference of 3.9 dB. For noise under the most sensitive conditions, the required 3^{rd} order resulted in a side lobe attenuation of around 23 dB, a widening of the beam of $10°$ or 14%, and a F/B-R_{25} difference of 0.3 dB. Speech at -10 dB and all position/orientations, as well as at 0 dB at P1 with $0°$ orientation, required an order of 1, resulting in a side lobe attenuation of 8 dB, a widening of the beam of $76°$ or 107%, and a F/B-R_{25} difference of around 10 dB. All other conditions with 0 dB direct sound did not require any modeling of the reference directivity except for diffuse-field equalization.

Tab. 2: Direct-to-reverberant energy ratio of the tested directivity patterns at the listener's ears in dB for all listen/source positions and orientation. Values that resulted in indistinguishable results for speech are printed bold.

directivity	0 dB direct sound			-10 dB direct sound			-20 dB direct sound		
order	P1, $0°$	P1, $180°$	P2, $0°$	P1, $0°$	P1, $180°$	P2, $0°$	P1, $0°$	P1, $180°$	P2, $0°$
7 (ref)	22.4	38.7	38.9	12.4	28.7	28.9	2.4	18.7	18.9
5	**22.5**	**38.7**	**38.9**	**12.5**	**28.7**	**28.9**	**2.5**	**18.7**	**18.9**
4	**22.7**	**38.7**	**38.9**	**12.7**	**28.7**	**28.9**	**2.7**	**18.7**	**18.9**
3	**22.9**	**37.9**	**38.7**	**12.9**	**27.9**	**28.7**	**2.9**	**17.9**	**18.7**
2	**24.5**	**36.1**	**39.0**	**14.5**	**26.1**	**29.0**	**4.5**	**16.1**	**19.0**
1	**27.8**	**33.9**	**40.8**	**17.8**	**23.9**	**30.8**	7.8	13.9	20.8
0	34.1	**34.1**	**42.4**	24.1	24.1	32.4	14.1	14.1	22.4

© Verband Deutscher Tonmeister e.V., 2019

Proceedings of ICSA 2019

5th International Conference on Spatial Audio
September 26th to 28th, 2019, Ilmenau, Germany

5. Conclusion

This paper evaluated the perceptual effect of reducing the spatial resolution (maximum spherical harmonics order) in directivity patterns for coincident source and listener position in a virtual room. For maximum sensitivity, the room simulation employed a higher-order image-source model without late diffuse reverberation and used dynamic binaural playback including head tracking that also controlled the orientation of the source. For the same reason, the reference directivity pattern was highly directive and the level of the direct sound was high, such as in human speech. The direct sound was played back omnidirectonal, i.e. inside the listener's head and the evaluation also included conditions with reduced direct sound to simulate other instruments.

In comparison to our previous experiment [7] with non-coincident listener/source positions, the perceptual influence of the reduction in spatial resolution was less pronounced, i.e. lower spherical harmonic orders were required to produce perceptually indistinguishable results from the reference. This could be attributed to the dominance of the direct sound in the new experiment. Thereby, reducing the direct sound increased the minimum required orders from 0 to 2, on average. This result agrees with the literature [16], where the sensitivity of the direct-to-reverberant energy ratio (DRR) is highest for values around 0 dB and decreases towards large absolute values of the DRR. Although the reduction of the spatial resolution yields an increase in beam width and reduction of side-lobe attenuation, the DRR is often well preserved, especially at the central listener/source position and direct sound levels as in human speech. In such cases, the diffuse-field equalization of the reduced-order directivity patterns might already be good enough. However, when facing a nearby wall and with less direct sound, the preservation of the directivity pattern is more important. The perceptual effect of the order reduction seems to be signal-dependent: coloration for noise, level and density of reverberation for speech.

Acknowledgments

This work is supported by the project Augmented Practice-Room (1023), which is funded by the local government of Styria via Zukunftsfonds Steiermark (future fond of Styria). The authors thank all listeners for their participation in the experiments and the reviewers for their helpful comments.

References

[1] B. N. J. Postma, H. Demontis, and B. F. G. Katz, "Subjective Evaluation of Dynamic Voice Directivity for Auralizations," *Acta Acustica united with Acustica*, vol. 103, no. 2, pp. 181–184, Mar. 2017.

[2] A. Kolarik, S. Cirstea, and S. Pardhan, "Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3395–3398, 2013.

[3] F. Wendt, F. Zotter, M. Frank, and R. Höldrich, "Auditory Distance Control Using a Variable-Directivity Loudspeaker," *MDPI Applied Science*, vol. 7, no. 7, 2017.

[4] F. Hohl, "Kugelmikrofonarray zur Abstrahlungsvermessung von Musikinstrumenten," Master's thesis, TU Graz, 2009.

[5] D. Deboy, "Tangential Intensity Algorithm for Acoustic Centering," in *Fortschritte der Akustik, DAGA*, Düsseldorf, 2011.

[6] I. B. Hagai, M. Pollow, M. Vorländer, and B. Rafaely, "Acoustic centering of sources measured by surrounding spherical microphone arrays," *Journal of the Acoustical Society of America (JASA)*, vol. 130, no. 4, 2011.

[7] M. Frank and M. Brandner, "Perceptual Evaluation of Spatial Resolution in Directivity Patterns," in *Fortschritte der Akustik, DAGA*, Rostock, Mar. 2019.

[8] C. Pörschmann, "One's Own Voice in Auditory Virtual Environments," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 378–388, 2001.

[9] J. S. Brereton, D. T. Murphy, and D. M. Howard, "The Virtual Singing Studio: A loudspeaker-based room acoustics simulation for real-time musical performance," in *Proceedings of the Baltic Nordic Acoustics Meeting (BNAM2012)*, 2012, pp. 18–20.

[10] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar. 2019.

[11] M. Romanov, P. Berghold, M. Frank, D. Rudrich, M. Zaunschirm, and F. Zotter, "Implementation and Evaluation of a Low-Cost Headtracker for Binaural Synthesis," in *Audio Engineering Society Convention 142*, May 2017.

[12] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik - DAGA*, Munich, March 2018.

[13] J. Daniel, "Représentation des champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2001.

[14] J. Pätynen and T. Lokki, "Directivities of Symphony Orchestra Instruments," *Acta Acustica united with Acustica*, vol. 96, no. 1, pp. 138–167, 2010.

[15] EBU, "EBU SQAM CD: Sound Quality Assessment Material recordings for subjective tests," 2008. [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[16] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008.

© Verband Deutscher Tonmeister e.V., 2019

View publication stats

4 Phonation Mode Analysis

The voice quality, a perceptual attribute partly defined by phonation type, of the singer is an important aspect in the analysis of the overall efficiency in singing. Automating the voice quality assessment could help improve voice education at the tertiary level and other fields of voice education. The analysis of phonation modes can provide information about expressed emotions or a certain singing style of a singer. These phonation modes then describe different oscillatory patterns of the vocal folds. The phonation modes studied in classical singing are not pathological. The transition from one to the other is more tenuous than the pathological phonation modes investigated in clinical studies. The signal processing used for the analysis can be parametric (glottal inverse filtering) or non-parametric and derived directly from a spectrogram (Mel-frequency cepstral coefficients). The performance of newly developed features derived from the modulation power spectrum is investigated in Section 4.1 and compared to commonly used reference features, which were derived from the aforementioned parametric and non-parametric approaches.

A software tool (VST plugin) for the analysis of phonation modes and vowel identification is presented in Section 4.2. Variants of several glottal inverse filtering techniques were studied in a preliminary investigation. We used the the most promising approach in the implementation of the VST plugin by enhancing classic linear prediction with homomorphic filtering. However, the results demonstrate that the considerable dependency on the fundamental frequency of the present glottal inverse filtering techniques limits their effectiveness for high pitched singing voices.

4.1 Classification of Phonation Modes in Classical Singing using Modulation Power Spectral Features

This work was published as:

M. Brandner, P. A. Bereuter, S. R. Kadiri, A. Sontacchi. Classification of Phonation Modes in Classical Singing using Modulation Power Spectral Features. *in IEEE Access*, 11:29149–29161, 2022. doi:10.1109/ACCESS.2023.3260187.

The idea and concept of this article were outlined by me and the second author. I wrote the original draft of the manuscript with periodical contributions from the second author, and the third author. The revision and editing was done by the second, third author and me. I conducted the measurements, the listening assessment and did all necessary preparations for the dataset. Most of the programming was done by me with help from the second author and with some contributions from the third author.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 February 2023, accepted 16 March 2023, date of publication 22 March 2023, date of current version 28 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260187

RESEARCH ARTICLE

Classification of Phonation Modes in Classical Singing Using Modulation Power Spectral Features

MANUEL BRANDNER[1], PAUL ARMIN BEREUTER[1], SUDARSANA REDDY KADIRI[2], (Member, IEEE), AND ALOIS SONTACCHI[1]

[1]Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, 8010 Graz, Austria
[2]Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland

Corresponding author: Manuel Brandner (brandner@iem.at)

This work was supported in part by the Academy of Finland under Project 330139.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Advisory Board of the University of Music and Performing Arts, Graz, and performed in line with the Declaration of Helsinki.

ABSTRACT In singing, the perceptual term ''voice quality'' is used to describe expressed emotions and singing styles. In voice physiology research, specific voice qualities are discussed using the term phonation modes and are directly related to the voicing produced by the vocal folds. The control and awareness of phonation modes is vital for professional singers to maintain a healthy voice. Most studies on phonation modes have investigated speech and have used glottal inverse filtering to compute features from an estimated excitation signal. The performance of this method is reported to decrease at high pitches, which limits its usability for the singing voice. To overcome this, this study proposes to use features derived from the modulation power spectrum for phonation mode classification in the singing voice. The exploration of the modulation power spectrum is motivated by the fact that, in singing, temporal modulations (known as vocal vibrato) and spectral modulations hold information of the vocal fold tension. Since there exists no large publicly available dataset of phonation modes in singing, we created a new dataset consisting of six female and four male classical singers, who sang five vowels at different pitches in three phonation modes (breathy, modal, and pressed). Experimental results with a support vector machine classifier reveal that the proposed features show better classification performance compared to state-of-the-art reference features. The performance for the current dataset is at least 10% higher compared to the performance of the reference features (such as glottal source features and MFCCs) in the case of target labels and around 6% higher in the case of perceptually assessed labels.

INDEX TERMS Modulation power spectrum, phonation modes, singing voice analysis, voice qualities.

I. INTRODUCTION

The classification of phonation modes as a computerised aid in classical singing voice training seems vital. Maintaining a healthy voice is an important component of professional singing and is essential for students during the course of their studies. The analysis and classification of phonation modes in classical singing might give a singer valuable insights into their voice production, and in the best case, prevent voice production problems. As voice production problems often occur throughout the course of voice studies, self-monitoring during vocal training and vocal warm-up could be beneficial to prevent more serious problems, which usually entail a longer rest period.

The phonation modes studied in classical singing are not pathological. The transition from one to the other is more tenuous than the pathological phonation modes investigated in clinical studies. Different phonation modes mean different

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

vibration patterns of the vocal folds, and in classical singing four main phonation modes can be distinguished. They are: *breathy*, *flow* (also referred to as resonant in [1]), *modal* and *pressed* phonation. *Breathy* phonation is characterized in [2] by minimal adductive tension, causing the vocal folds to reside in a Y-shaped state which leaves an opening at the top of the vocal folds, even during the closure phase of one glottal cycle. This constant opening lets the turbulent tracheal airflow enter the vocal tract at any time causing a *breathy* voice perception, which is also referred to as aspiration noise [3]. Physiologically, *modal* phonation in singing voice is defined by a full-length vibration of the vocal folds caused by moderate tension and compression, which implicitly leads to a full glottal closure of the vocal folds during vibration [2], [4]. The opposite end of the phonatory dimension described in [2] is given with *pressed* voice, which is defined by strong subglottic pressure and adduction caused by tense muscles surrounding the vocal folds.

The differences between the phonation modes have been studied for speech, using features derived by signal processing methods that attempt to separate vocal fold movement information (excitation signal or glottal source waveform) from the vocal tract contribution (filter). A common approach of calculating glottal waveform characteristics is by using a glottal inverse filtering method (GIF) (based on *source-filter deconvolution*) [5], [6], [7], [8], [9]. An overview of glottal source processing is given in [10]. Although speech and singing share similar basic concepts of voice production, the analysis of singing voices is far more difficult from a signal processing point of view, especially due to higher and rapid variations in pitch. As the pitch increases, the analyzed signal in the frequency domain exhibits an increased sparsity, especially for sustained vowels due to its harmonic structure, which leads to an ill-posed mathematical condition for GIF methods. Automated inverse filtering usually fails at higher pitches and some implementations are reported to be already erroneous at above ca. 300 Hz [11]. Features used for phonation mode classification, which are derived from an estimated glottal waveform using a GIF-method are commonly denoted as voice quality features (VQ-features) [12], [13].

Another common parametric method for separating the vocal tract contribution and calculating glottal source characteristics is cepstral analysis [14], which was initially introduced in [15] for seismic analysis in order to find echo components. The difference between cepstral analysis and glottal inverse filtering lies in the fact that the information of both the source and the filter are found in the same resulting cepstral domain signal, but at different locations along the so-called quefrency axis. The magnitude of the first peak along the quefrency axis has been found to be well suited to determine the breathiness of the voice [16] and, when used as a feature, is called cepstral peak prominence (CPP). In [17], it is concluded that CPP is similar to the first rahmonic and gives meaningful results to detect breathiness. In [18] CPP was reported to separate neutral, breathy, and pressed phonation from each other, but not flow from the other phonation modes. Also, breathy phonation was shown to have high level of turbulent noise and is reported to have a large harmonic to noise ratio (HNR) [19]. Modal phonation in singing results in rich harmonics and pressed phonation is reported to typically show a weaker fundamental and more dominating higher harmonics [20]. Mel-frequency cepstral coefficients (MFCCs) [21] serve as a descriptive representation of the magnitude spectrum and are frequently used for classification of phonation modes in [13] and [18].

Unlike the cepstral analysis, the modulation spectrum comprises temporal information and results from an analysis along the temporal axis and not along the frequency axis. In [22], the significance of low-frequency modulations is discussed along with how to use the modulation spectrum to analyze sound in accordance with the human auditory system. Studies in [23] and [24] showed that temporal modulations influence speech intelligibility. However, the extraction of characteristics from the modulation spectrum is not trivial, due to its high dimensionality. This is why in [25], [26], and [27] it is proposed to apply a Higher Order Singular Value Decomposition (HOSVD) on the modulation spectrum in combination with a feature selection algorithm based on the mutual information. The results for their approach for voice pathology detection achieves a detection rate of around 94% and to classify hoarseness, a global classification rate of 74% is reported. However, they did not use the modulation power spectrum as originally presented in [28] and [29], which combines the advantages of cepstral analysis and the modulation spectrum by extracting both temporal and spectral modulations. Moreover, they limited their investigations to a single vowel (/a/). The most comparative studies investigating phonation modes in classical singing have been [13], [18], and [1], but unfortunately all studies used a small dataset with data of only two singers. According to the authors' knowledge, there are no studies on classification of phonation modes in classical singing using characteristics extracted from the modulation power spectrum and no previous work has investigated phonation modes on a larger dataset consisting of data from more than two classical singers.

II. GOALS OF THE CURRENT STUDY

In classical singing, vocal vibrato is a temporal modulation which lies at 4 to 8 Hz [30], [31] and spectral modulations depend on the spectral composition of a sung vowel, which provide information on the harmonic structure of a sound. The spectral composition can show how breathy or strained the voice sounds, which depends on the singer's physical effort on the vocal folds and the amount of used airflow [32]. In order to investigate both the temporal and spectral modulations of sung vowels, we propose the investigation of novel features extracted from the modulation power spectrum (MPS) [28], [29]. This method combines the benefits of the discussed parametric and non-parametric approaches. We investigate a peak-picking technique similar to CPP, where we additionally include higher harmonics along the temporal and spectral modulation axes, as opposed

to an algebraic approach like the HOSVD utilized in [25], [26], and [27]. As of now, the data made public in [1] and [20] are the only two openly accessible datasets of professional singers, singing with different phonation modes. However, both of them have severe restrictions regarding the number of singers and ratings. Thus, we have created a new dataset including ten singers singing five vowels in three phonation modes (breathy, modal, and pressed) over a large pitch range. We propose two novel feature sets derived from the modulation power spectrum and one feature set derived from an averaged cepstrum over consecutive time frames for the classification of the phonation modes *breathy*, *modal* and *pressed*. The proposed feature sets are compared to three state-of-the-art reference feature sets. The feature sets are compared by means of their classification performance using a support vector machine (SVM) classifier (see section V).

The highlights and novelties of the current study are:

- Investigation of temporal and spectral characteristics extracted from the modulation power spectrum.
- Investigation of automatic classification of phonation modes (breathy, modal, and pressed) on a newly created classical singer dataset.
- Investigation of a feature reduction by using the averaged MPS along the temporal axis.
- Study of the performance of features derived from the averaged cepstrum over consecutive time frames compared to MPS features.
- Comparison of the proposed features with state-of-the-art reference features, which are: voice quality features (VQ-features), cepstral features (MFCCs), and features derived after zero frequency filtering (ZFF features).

The organization of the paper is as follows: Section III describes the data collection including measurements and labelling. The extraction of features and calculation of the modulation power spectrum are described in section IV. The experimental protocol is described in section V, which gives a general overview of the classification framework, the reference features, the classifier, and the evaluation metrics. Results of the classification experiments are presented in section VI. In section VII the results are discussed and section VIII summarizes the study.

III. DATA COLLECTION

For the present work, we created a new dataset of audio recordings of sustained vowels sung at various pitches with three different phonation modes. Furthermore, we conducted a listening assessment in order to obtain the perceived phonation modes of the recorded vowels. Although datasets on phonation modes in singing exist, the current dataset is much larger compared to existing datasets [1], [20]. The already available datasets hold recordings of only two classical singers, whereas the proposed dataset contains recordings of ten classical singers. In contrast to the already available data, where the labels of phonation modes are based solely on the judgements of a single expert, a perceptual assessment was performed for the presented dataset, resulting in 6 ratings per recording. This allows for a statistical analysis of the perceived phonation mode labels. The dataset is named as Voice Qualities in Singing (VQS) and is publicly available at: https://phaidra.kug.ac.at/o:126552.

A. MEASUREMENTS

The measurements were recorded with a microphone (omnidirectional pattern, NTI M2230, Schaan, Liechtenstein) at a distance of 1 m in front of the singer. For the acoustic analysis, dry signals measured in an anechoic environment are ideal. However, in singing, room acoustics support the voice, which is a necessity in a longer recording session. Therefore, we used an augmented acoustic system with zero latency [33], [34], which only gives the singer natural room acoustics via transparent headphones [35] while creating no reverberation on the microphone signals. The augmented acoustic system is fed with the signal of the microphone placed in front of the singer and employs a static, however frequency-dependent directivity to excite the virtual room. The virtual room simulates a shoe-box-like concert hall with a size of roughly 30 m × 24 m × 20 m and reverberation time of 2.2 s. Typical reverberation times of concert halls are in the range between 1.5 s and 3 s [36], [37].

1) ROOM CONDITIONS

Measurements were carried out in a sound treated measurement room with absorptive material on the walls and floor at the Institute of Electronic Music and Acoustics in Graz. The frequency-dependent room reverberation in the measurement room is less than 75 ms between 400 Hz and 1 kHz, and less than 50 ms above 1 kHz. The volume of the room is approximately 50 m^3 with a floor space of 22.50 m^2.

B. AUDIO RECORDINGS

Four male singers (3 tenors and 1 baritone) were instructed to sing 5 sustained German vowels (/a:/, /e:/, /i:/, /o:/, and /u:/) over the pitch range from H/B2/123 Hz to a^1/A4/440 Hz on a whole-tone scale, except for the baritone, who only sang up to e^1/E4/330 Hz. Furthermore, six female singers (3 sopranos and 3 mezzo-sopranos) sang the vowels from a/A3/220 Hz to a^2/A5/880 Hz. The singers were asked to sing the vowels, starting on the consonant /m/ and sustaining the vowel for 2 seconds. The vowels were repeated three times each with different provoked voice phonation modes (modal, breathy, and pressed). This results in a total number of 2145 audio samples. The dataset was then reduced to a total number of 1140 samples to ensure a listening assessment of reasonable duration (see section III-C). This reduced dataset is consequently used in the classification experiments presented in section V.

All singers were trained classical singers except for the baritone (studied jazz vocals), who said to have the ability to mimic the classical singing technique due to his teaching experience at the music conservatory. The average age was 29.6 years (the youngest was 24 years and the oldest

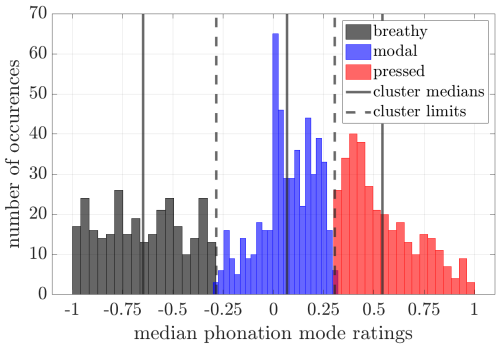


FIGURE 1. Histogram of perceptually assessed median phonation mode ratings along with the cluster limits (which are used to derive the perceptually assessed labels).

was 34 years). The classically trained singers were 4 graduate students (at the end of their current master studies), 5 post-graduate students (with one master's degree or more), and 1 undergraduate student (bachelor's degree). Six of the singers were also teaching at the time. The singers were asked to sing at a comfortable loudness level (mezzo-forte). All the participants were well-trained for the task due to their extensive practice during their classical vocal studies.

a: NOTE ON THE PHONATION MODES

We chose to study three phonation modes: breathy, modal, and pressed, as most singers were unfamiliar with the term ''flow'' phonation. In this sense, the term ''modal'' in our study and in the dataset defines the optimal singing voice phonation, and the other two phonation modes ''breathy'' and ''pressed'' are deviations from this optimal state.

C. TARGET AND PERCEPTUALLY ASSESSED LABELS

The instructions given to the singers during the course of the recordings are used as target labels in this work. Due to the time-consuming nature of a listening assessment, only a portion of the recordings were chosen to be perceptually assessed. This resulted in the selection of 1140 samples. The pitches of these samples are listed in Table 1. The samples were randomly grouped into ten subsets, which each were independently rated by 6 listeners, resulting in six independent ratings per sample. A total of 20 listeners participated in the assessment. As a starting point for the current investigation, a k-mediods clustering algorithm is used to categorize the median of the six independent ratings for each sample. The distributions of the three phonation mode clusters derived using the k-mediods algorithm are visualized in Fig. 1 as histograms. The cluster boundaries are chosen as the upper and lower boundary of the modal cluster, which provides one possible straight-forward approach for the assignment of a fixed label to each recording. Fig. 1 also shows a smaller distance between the cluster medians of the modal and pressed cluster compared to the distance between the modal and breathy cluster. From the resulting 1140 samples, 297 samples were rated as breathy, 516 as modal, and 327 as pressed. The amount of data for target labels and perceptually assessed labels are listed in Table 2 along with information of gender, vowels, and pitch range. The confusion matrix in Table 3 presents the differences between the perceptually assessed and the target labels. When comparing the different labels of phonation mode classification, it can be seen that pressed and modal phonation consistently cause the greatest uncertainty, suggesting that performance differences in classifications are to be expected. However, it is also reasonable to anticipate that there will be some uncertainty in the data, if only target labels (instructions to the singers) are looked into.

IV. FEATURE EXTRACTION

In this section, newly developed features, based on the modulation power spectrum (MPS) including their underlying theory, are discussed along with the features derived from the averaged cepstrum over consecutive time frames. In order to calculate the MPS-based features, a peak-picking procedure is applied, which uses the knowledge of the fundamental frequency. Three sets of features are proposed, one set is based on the two-dimensional MPS-representation, leading to a larger dimension feature set (MPS_{peaks}), the second set builds on a compact, summed version of the MPS resulting in a smaller dimension feature set (MPS_{sum}), and the third set is derived based on cepstral peaks ($Ceps_{peaks}$), as there exists a strong relation between the cepstrum and the spectral modulation dimension of the MPS. A schematic block diagram describing the steps involved in the computation of the three feature sets is shown in Fig. 2.

A. MODULATION POWER SPECTRUM

The origin of the modulation power spectrum (MPS) can be traced back to the field of neuroscience, where it was used to better understand human auditory processing [28]. Temporal modulations which constitute the modulation spectrum along the time axis, have been studied for different tasks such as audio coding, modification, and automatic classification [22]. Subsequently, temporal modulations have also been extensively investigated for pathological voices [25], [26], [27]. The MPS combines the information of temporal modulations and the approach of cepstral analysis, which aims at a separation of vocal tract and voice source information [14]. The MPS is calculated by applying a two-dimensional Fourier transform on the squared and logarithmized amplitude values of a short-time Fourier transform (STFT) ($X(m,k)$), computed with the block-length L and hop-size R. The STFT consists of N positive frequencies and M time-frames, where the current time-block is denoted using the index m and k denotes the discrete frequency indices. Note that instead of using the natural logarithm as mentioned in [28], we use the logarithm with base 10, and represent the spectro-temporal modulation amplitudes $S(k_f, k_t)$ in decibels.

$$S(k_f, k_t) = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} 10 \log_{10}(\|X(m,k)\|^2) e^{-j2\pi U}, \tag{1}$$

TABLE 1. Pitches of the data selected from the full dataset which were perceptually assessed. Frequency differences in Hz between the pitches (delta) are also listed to show the almost linear spacing.

pitch	$c/C3$	$g/G3$	$c^1/C4$	$e^1/E4$	$g^1/G4$	$a^1/A4$	$c^2/C5$	$d^2/D5$	$e^2/E5$	$g^2/G5$	$a^2/A5$
Hz	131	196	262	330	392	440	523	587	659	788	880
delta	-	65	66	68	62	48	83	64	72	129	92

TABLE 2. Overview of the analyzed dataset* with information on gender, vowels, pitch range, and the number of target and perceptually assessed phonation mode labels.

gender	female singers (6)	male singers (4)
vowels	/a:, e:, i:, o:, u:/	/a:, e:, i:, o:, u:/
pitch range	$c^1/C4/261Hz$ to $a^2/A5/880Hz$	$c/C3/131Hz$ to $a^1/A4/440Hz$ **
target	breathy (270), modal (270), pressed (270)	breathy (110), modal (110), pressed (110)
assessed	breathy (221), modal (359), pressed (230)	breathy (76), modal (157), pressed (97)

* Subset of the full dataset at selected pitches.
** The baritone only sung up to e^1/E4/330 Hz.

TABLE 3. Percentage of confusions: *target labels* (rows) vs. *perceptually assessed labels* (columns). The perceptual listening assessment's results are compared with the reference number of target labels per class (380).

	Breathy [%]	**Modal** [%]	**Pressed** [%]
Breathy	74	20	6
Modal	4	71	25
Pressed	1	45	54

where $U = \left(\frac{mk_t}{M} + \frac{k_f k}{N}\right)$. The indices $k_f = -\left\lfloor \frac{N}{2} \right\rceil, \ldots, \left\lfloor \frac{N}{2} \right\rceil$ and $k_t = -\left\lfloor \frac{M}{2} \right\rceil, \ldots, -\left\lfloor \frac{M}{2} \right\rceil$[1] are the corresponding discrete spectral and temporal modulation frequency bins in the joint modulation frequency domain after the two-dimensional Fourier transform. In our implementations, a Blackman-Harris window with a block-length of 80 ms, a hop-size of 2.5 ms, and a 4096-point fast Fourier transform at a sampling frequency of 16 kHz are used. The block length is longer than commonly used in speech signal processing (25 to 50 ms) to accommodate the nature of the singing voice. The most important aspect is the choice of a block length and a hop size that allow to study the vibrato characteristics of the classical singing voice, which has a vibrato frequency around 4 to 8 Hz [30].

B. EXTRACTION OF MODULATION POWER SPECTRAL FEATURES

The MPS represents a high dimensional feature space and exhibits pitch-dependent regions of high and low spectro-temporal energy, especially in the case of sustained vowels. Therefore, we propose a pitch-normalized peak-picking strategy to extract only the high energy components of the MPS. Fig. 3 shows illustrations of modulation power spectra for three phonation modes (breathy, modal, and pressed). The search regions depicted in Fig. 3 on the spectral modulation axis (y-axis) with a width of $\pm\frac{1}{3}\tau_0$ are centered around $n_f \cdot \tau_0$, with $n_f = \{1, 2, \ldots, N_f = 8\}$ being positive multiples of the fundamental period $\tau_0 = \frac{1}{f_0}$. The fundamental periods are determined using the reference pitches listed in Table 1. The search regions on the temporal modulation axis are fixed around $N_t = 5$ multiples, centered at $n_t = \{-2, -1, 0, 1, 2\}$, times a pre-selected vibrato frequency of $f_{vib} = 6$ Hz.[2] The search region width along the x-axis is chosen to be $\pm\frac{1}{3}f_{vib}$. The boundaries of the spectral and temporal search regions are formulated in (2) and (3).

$$\tau_i = n_{f,i} \cdot \tau_0 \pm \frac{1}{3}\tau_0 \tag{2}$$

$$f_{t_{mod},i} = n_{t,i} \cdot f_{vib} \pm \frac{1}{3}f_{vib} \tag{3}$$

The modulation frequencies $f_{t_{mod}}$ and τ denote the temporal and spectral frequencies used in the modulation power spectrum. They can be calculated using the linear relationship between the discrete modulation frequency bins k_t and k_f and the spectral modulation frequency resolution Δ_τ and temporal modulation frequency resolution $\Delta_{f_{t_{mod}}}$ (see (4) and (5)).

$$\tau = \Delta_\tau \cdot k_f \tag{4}$$

$$f_{t_{mod}} = \Delta_{f_{t_{mod}}} \cdot k_t \tag{5}$$

These modulation frequencies are used below to describe the process of calculating the newly proposed features from the modulation power spectrum.

MPS_{peaks} : The full set of peak amplitudes derived as in (6) is denoted as MPS_{peaks}.

$$MPS_{peaks} = \max\{S(\tau_i, f_{t_{mod},i})\} \tag{6}$$

The dimension of this feature set is: $N_t \cdot N_f = 5 \cdot 8 = 40$ peak amplitudes.

MPS_{sum} : In order to further reduce the dimensionality of the full set of peak values MPS_{peaks}, the MPS is summed along the temporal modulation axis, see (7).

$$S_\Sigma(k_f) = \sum_{k_t} S(k_t, k_f) \tag{7}$$

[1] $\lfloor \cdot \rceil$ denotes a rounding operation.

[2] f_{vib} is chosen as the mean value of the vibrato range 4 to 8 Hz reported in [30].

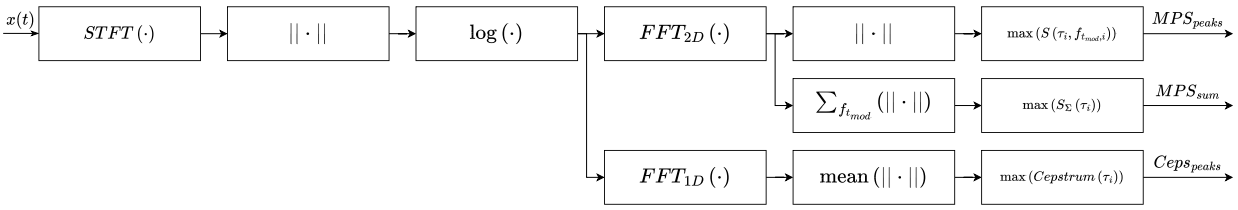
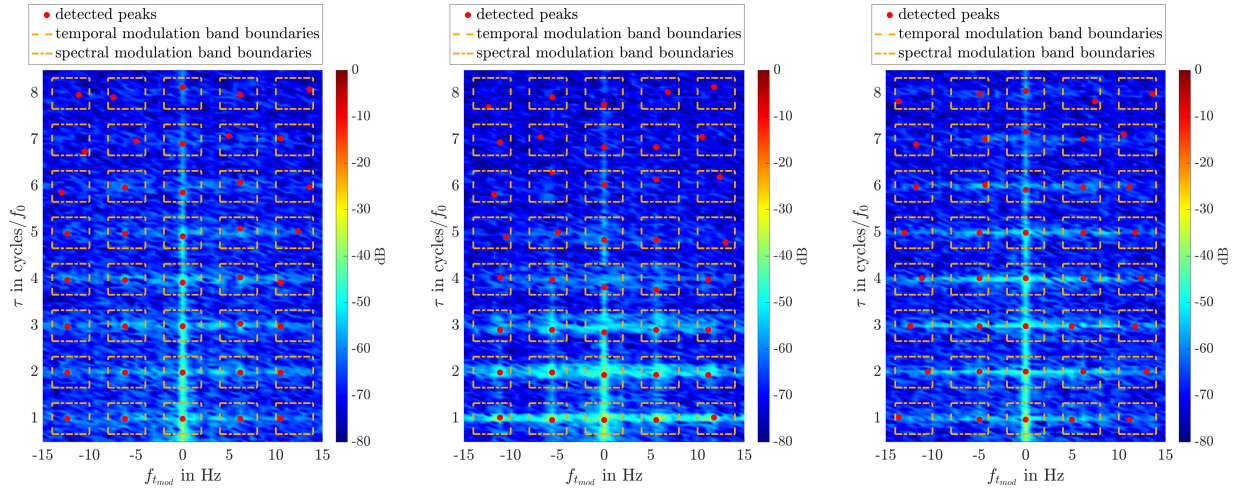


FIGURE 2. Schematic block diagram for the extraction of features from the modulation power spectrum (MPS_{peaks} and MPS_{sum}) and the averaged cepstrum ($Ceps_{peaks}$).



(a) breathy. (b) modal. (c) pressed.

FIGURE 3. Illustrations of modulation power spectra for the three phonation modes. Shown are the extracted peak amplitudes within a search grid referenced to the fundamental frequency ($\tau_0 = \frac{1}{f_0}$) and the average temporal modulation frequency for vibrato $f_{vib} = 6\,\text{Hz}$ along both the temporal and spectral axes. The search grid regions along the spectral axis lie within $n_f \cdot \tau_0 \pm \frac{1}{3}\tau_0$ with $n_f \in \{1, 2, \ldots, 8\}$, and along the temporal axis within $n_t \cdot f_{vib} \pm \frac{1}{3} f_{vib}$ with $n_t \in \{-2, -1, 0, 1, 2\}$.

Again, 8 peaks along the spectral modulation axis are picked (see (8)), which is the dimension of the MPS_{sum} feature set.

$$MPS_{sum} = \max\{S_\Sigma(\tau_i)\} \tag{8}$$

Fig. 3 shows the modulation power spectra for the breathy, modal, and pressed phonation modes. The MPS for breathy phonation shown in Fig. 3(a), is characterized by low energy along the temporal modulation axis, and evenly distributed energy along the spectral modulation axis around $f_{t_{mod}} = 0$ Hz. Fig. 3(b) shows the illustration of the MPS for modal phonation, exhibiting strong temporal modulation components, but also a substantial energy decrease along the spectral modulation axis. The MPS for pressed phonation shown in Fig. 3(c) shows more energy at higher spectral modulation frequencies. Fig. 4 shows the illustrations of the summed modulation power spectra for the three phonation modes. From the illustrations of the modulation power spectra (shown in Fig. 3 and Fig. 4), it is clearly evident that temporal modulation components and spectral harmonic structures vary among the three phonation modes.

C. EXTRACTION OF FEATURES FROM THE AVERAGED CEPSTRUM

There exists a strong relationship between the spectral modulation dimension of the MPS and the cepstrum. Therefore, peak values extracted from the averaged cepstrum over consecutive time frames are also considered as possible features. The calculation of the cepstrum is presented in Fig. 2, whereas the features extracted from it are denoted as $Ceps_{peaks}$. We extracted the peak amplitudes within the averaged cepstrum using the same search regions as for the spectral modulation axis in the MPS (see Sec. IV-B).

V. EXPERIMENTAL PROTOCOL

In order to investigate the phonation mode classification performance of the newly proposed modulation power spectral features we set up a classification problem. The newly developed features are compared with state-of-the-art reference features that are commonly used and have been employed in previous comparative work [1], [13], [18]. In the current study, we use a support vector machine (SVM) including a hyperparameter optimization, and a leave-one-singer-out

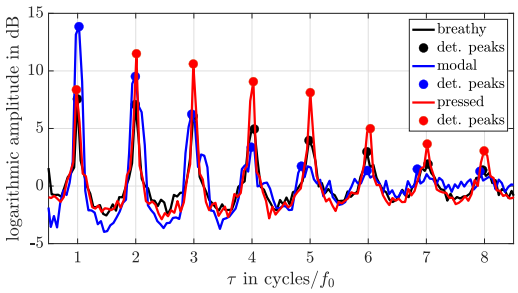


FIGURE 4. Illustrations of the summed modulation power spectra for the three phonation modes. The circular markers indicate the peak extraction for the corresponding MPS_{sum} feature. The summed modulation power spectra are detrended by subtracting a fitted first-order polynomial for better visualization.

(LOSO) cross-validation technique. This may reduce overall performance accuracy, but should increase the generalizability of the current results. The basic processing and classification framework is shown as a block diagram in Fig. 5. The feature extraction block is preceded by the pre-processing steps, which includes segmenting the audio samples to the sustained part of the vowels and excluding the consonant /m/ at the beginning. The feature extraction block summarizes the computation of the reference features, as well as the steps for calculating the newly developed features based on the MPS and the averaged cepstrum (averaged over consecutive time frames, see section IV). The last block indicates the classification process with the SVM, which predicts one of the three phonation modes for each audio sample after the classifier has been trained according to the LOSO cross-validation and hyperparameter optimization technique. The following sections present the reference features, the details of the classifier and the evaluation metrics along with the classification framework.

A. REFERENCE FEATURES

The proposed features extracted from the modulation power spectrum are compared to three state-of-the-art reference feature sets (see Fig. 5), which are briefly described in the following paragraphs.

1) VOICE QUALITY FEATURES (VQ)

The VQ feature set consists of six features, derived from a glottal waveform estimate, which is calculated using a GIF-method [12], [13]. The six features are: (1) normalized amplitude quotient (NAQ) [38], (2) quasi-open quotient (QOQ) [10], [39], (3) amplitude difference between fundamental and first harmonic (H1-H2) [39], (4) parabolic spectral parameter (PSP) [40], (5) harmonic richness factor (HRF) [39] and (6) maximum dispersion quotient (MDQ) [12]. The literature shows that voice quality features work well for speech, but their applicability to singing is known to be limited at high pitches, due to erroneous glottal inverse filtering [11]. Nevertheless, we use these features as reference features in the current study.

2) ZERO FREQUENCY FILTERING (ZFF)

The ZFF method provides an approximate voice source waveform without explicitly using the source-filter model of speech production. The ZFF feature set consists of four features, which are: the strength of excitation (SoE), the energy of excitation (EoE), the loudness measure and the ZFF signal energy. These features were shown to be useful for discriminating phonation types in speech and singing [13], [41], [42]. SoE was shown to be proportional to the rate of glottal closure, the EoE feature was shown to capture the vocal effort, and the loudness measure was shown to capture the abruptness of the glottal closure [43], [44]. The energy of the ZFF signal at glottal closure is also used as a feature which was shown to capture low frequency energy [13]. Zero frequency filtering features have been designed to overcome the problem of the classical voice quality features and have been extensively investigated in [13], but it has been shown that the performance for singing voices could still not be improved significantly.

3) MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs)

MFCCs are popular features used in many tasks, such as automatic speech recognition [45], [46], music information retrieval [47], [48], including phonation modes classification in speech and singing [18]. The MFCCs are derived using the same parameters as for the MPS (Blackman-Harris window, 80 ms window length and 2.5 ms hop-size). From the mel-cepstrum, the first 36 cepstral coefficients are derived. The 0^{th} coefficient is not considered, which results in a 35-dimensional feature vector. MFCCs have been shown to be versatile descriptors for speech recognition tasks and phonation mode classification in several previous works [13], [18]. In comparison to other features, MFCCs are harder to interpret and their descriptive quality usually depends largely on the number of used coefficients.

B. FEATURE SET COMBINATIONS

Additionally, the reference and proposed features described above were combined in order to study the complementary information among the features. In total, 9 combinations of feature sets (FSCs) were created in addition to the single feature sets and their corresponding performances are discussed in subsection VI-C.

C. CLASSIFIER

We use a SVM with a radial basis function kernel as a classifier [49], because it is reported to perform well even on a smaller number of training data. We perform a hyperparameter tuning with GridSearchCV [49] within a LOSO cross-validation strategy to avoid over-fitting and increase the generalizability of the model. For our phonation modes classification task, we use two types of labels: (i) the instructed phonation modes given to the singers during recordings (target labels) and (ii) the perceptually asssessed phonation mode labels from the listening assessment (perceptually

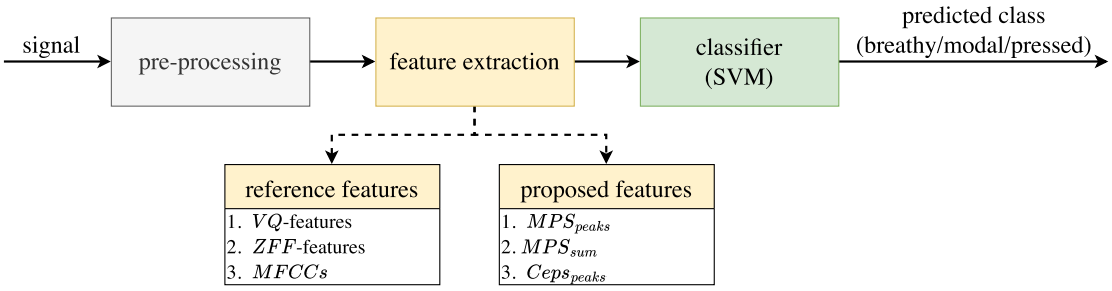


FIGURE 5. Block diagram of the basic processing and classification framework including the pre-processing stage, the feature extraction for the reference and proposed features, and the classification using a support vector machine to predict one of the three phonation modes for an audio sample. The depicted framework indicates the processing after the SVM classifier has been trained according to the LOSO cross-validation and hyperparameter optimization technique.

assessed labels). Experiments are conducted by considering both genders (including both male and female singers, totally 10 singers) and only female singers (6 singers).

D. EVALUATION METRICS

Performance measures are the mean and standard deviation of the test accuracy over the runs of the LOSO cross-validation (for the whole dataset and the female-only dataset). We omit the male-only dataset due to its small sample size. We have computed the standard deviation of the accuracy for each feature set to see the reliability of the features across varying singers. As an additional metric, we provide confusion matrices for the test sets averaged over all runs of the LOSO cross-validation to examine the confusions among phonation modes.

VI. RESULTS

In this section, the results of the classification problem described in section V are presented for the target and perceptually assessed labels, in terms of accuracy and confusion matrices (see sections VI-A and VI-B). In each of the subsections, we present the results (for both the reference and proposed features) for the whole dataset, i.e., combination of male and female singers (see Table 2 for an overview of the data and labels) and the female-only dataset. Finally, the results of the feature set combinations are listed in section VI-C.

A. CLASSIFICATION RESULTS FOR TARGET LABELS

This section presents the classification results (in terms of mean and standard deviation of accuracies) obtained for the target labels. The classification accuracies for the whole data are given in Table 4. From the table, it is observed that the proposed feature sets (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) perform better than the reference features. All the proposed features show a mean accuracy which is around 10% higher than the mean accuracy of the reference features. However, the standard deviations of the accuracies are all larger for the proposed features. The most striking aspect in Table 4 is the similar performance of the 40-dimensional MPS_{peaks} feature set and the 8-dimensional MPS_{sum} and $Ceps_{peaks}$ feature sets. In order to gain more information on the misclassifications, confusion matrices are given in Table 5 and Table 6 for one proposed feature set (MPS_{sum}) and one reference feature set (VQ-features). The confusion matrices clearly show that there exists greater confusion between pressed and modal, and between breathy and modal phonation modes in the VQ feature set, compared to the proposed MPS_{sum} feature set.

The results for the female-only data are given in Table 7. It is expected that the female-only data, still consisting of 810 samples, will be more homogeneous because all singers

TABLE 4. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the classical singers (female+male) using the *target labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference features:		
MFCCs (35)	57	± 9
VQ (6)	46	± 6
ZFF (4)	51	± 6
Proposed features:		
MPS_{peaks} (40)	**68**	± **14**
MPS_{sum} (8)	67	± 12
$Ceps_{peaks}$ (8)	66	± 11

TABLE 5. Confusion matrix in % for the MPS_{sum} feature set of the whole dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	77	12	11
Modal	8	76	16
Pressed	12	32	56

TABLE 6. Confusion matrix in % for the VQ feature set of the whole dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	60	29	11
Modal	27	47	26
Pressed	23	45	32

TABLE 7. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the female classical singers using the *target labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference Features:		
MFCCs (35)	61	$\pm$ 11
VQ (6)	48	$\pm$ 4
ZFF (4)	56	$\pm$ 3
Proposed Features:		
MPS_{peaks} (40)	74	$\pm$ 9
MPS_{sum} (8)	**75**	$\pm$ **7**
$Ceps_{peaks}$ (8)	68	$\pm$ 10

TABLE 8. Confusion matrix in % for the MPS_{sum} feature set of the female-only dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	88	5	7
Modal	6	74	20
Pressed	9	27	64

TABLE 9. Confusion matrix in % for the ZFF feature set of the female-only dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	64	24	12
Modal	30	44	26
Pressed	12	28	60

sang the same pitches. Similar to the results of the whole dataset, it can also be observed that the proposed feature sets (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) perform better than the reference features. Moreover, it can be seen that the results show lower standard deviations compared to the results for the whole dataset, except for the MFCCs. The classification accuracies are also increased for all the feature sets by 4-8% for the female-only data, where the MPS_{sum} feature set showed the highest accuracy among all features, and the lowest standard deviation among the proposed feature sets. The confusion matrix shown in Table 8 demonstrates the increased performance for the MPS_{sum} feature set, concerning less confusion between breathy and pressed phonation modes compared to the results for the whole dataset, but no performance increase for the classification of modal phonation. On the other hand, the confusion matrix for the ZFF feature set given in Table 9 indicates that there exists greater confusion with modal for breathy and pressed phonation modes compared to the MPS_{sum} feature set in the female-only data.

B. CLASSIFICATION RESULTS FOR PERCEPTUALLY ASSESSED LABELS

This section gives the classification results obtained for the perceptually assessed labels. The classification accuracies for the whole data are given in Table 10. The table, shows that the proposed feature sets (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) again perform better than the reference features. In general, the performance differences between the feature sets decrease when perceptually assessed labels are used.

The results for the female-only data evaluated for the perceptually assessed labels are given in Table 11. Again,

TABLE 10. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the classical singers (female+male) using the *perceptually assessed labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference Features:		
MFCCs (35)	59	$\mp$ 7
VQ (6)	52	$\mp$ 9
ZFF (4)	54	$\mp$ 9
Proposed Features:		
MPS_{peaks} (40)	64	$\mp$ 6
MPS_{sum} (8)	**65**	$\mp$ **8**
$Ceps_{peaks}$ (8)	61	$\mp$ 6

TABLE 11. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the female classical singers using the *perceptually assessed labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference Features:		
MFCCs (35)	65	$\mp$ 3
VQ (6)	53	$\mp$ 5
ZFF (4)	55	$\mp$ 4
Proposed Features:		
MPS_{peaks} (40)	64	$\mp$ 6
MPS_{sum} (8)	**69**	$\mp$ **5**
$Ceps_{peaks}$ (8)	64	$\mp$ 5

the highest mean accuracies are obtained for the MPS_{sum}, followed by the MPS_{peaks}, $Ceps_{peaks}$ and MFCC feature set.

Overall, the results in the classification experiments (both for target and perceptually assessed labels) show that the features extracted from the modulation power spectrum (MPS_{peaks}, and MPS_{sum}), and the cepstral features ($Ceps_{peaks}$) perform better than the reference features. Most striking is the performance of the MPS_{sum} feature set, which only consists of 8 features. This suggests that including modulation characteristics in phonation mode analysis is beneficial, especially in the analysis of classical singing.

C. CLASSIFICATION RESULTS FOR COMBINATIONS OF FEATURE SETS

This section reports the results for the combination of the proposed and the reference features on the whole dataset for the target and perceptually assessed labels. The feature sets are combined to investigate the complementary information among the feature sets. In total, 9 feature set combinations (FSC) for each label group (target and perceptually assessed labels) were created as listed in Table 12. FSC1 and FSC2 include combinations of the reference feature sets. FSC3 to FSC5 combine the best reference feature set combination with the proposed feature sets. FSC6 to FSC8 combines the proposed feature sets to investigate their corresponding complementary information. FSC9 combines the best performing feature sets of the reference feature set combinations and the proposed feature set combinations. The feature set combination FSC8 produces the highest mean accuracies for the target labels and FSC9 for the assessed labels. Interestingly, the combination of the newly proposed features FSC8 (MPS_{peaks} and $Ceps_{peaks}$) and FSC7 (MPS_{sum} and $Ceps_{peaks}$) perform

TABLE 12. Feature set combinations FSC1 to FSC9 and their corresponding accuracies (mean and standard deviation in percent) for the target and perceptually assessed labels.

	target (f.+m.)			percep.ass.(f.+m.)		
ID	**combined features**	**Mean acc.** [%]	**Std.** [%]	**combined features**	**Mean acc.** [%]	**Std.** [%]
FSC1	VQ+ZFF	51	7	VQ+ZFF	58	6
FSC2	ZFF+MFCCs	59	11	VQ+ZFF+MFCCs	65	8
FSC3	ZFF+MFCCs+MPS_{peaks}	69	14	VQ+ZFF+MFCCs+MPS_{peaks}	67	7
FSC4	ZFF+MFCCs+MPS_{sum}	67	14	VQ+ZFF+MFCCs+MPS_{sum}	68	7
FSC5	ZFF+MFCCs+$Ceps_{peaks}$	66	14	VQ+ZFF+MFCCs+$Ceps_{peaks}$	67	11
FSC6	MPS_{peaks}+MPS_{sum}	68	15	MPS_{peaks}+MPS_{sum}	67	7
FSC7	MPS_{sum}+$Ceps_{peaks}$	68	13	MPS_{sum}+$Ceps_{peaks}$	68	7
FSC8	MPS_{peaks}+$Ceps_{peaks}$	**70**	**14**	MPS_{peaks}+$Ceps_{peaks}$	65	5
FSC9	ZFF+MFFCs+MPS_{peaks}+$Ceps_{peaks}$	70	15	VQ+ZFF+MFCCs+MPS_{sum}+$Ceps_{peaks}$	**70**	**7**

TABLE 13. Confusion matrices for all the reference features (MFCC, VQ, and ZFF) and proposed features (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) of the whole dataset (combination of female and male data) and the female-only dataset using the target labels and perceptually assessed labels. Here B, M and P refer to breathy, modal, and pressed phonation modes, respectively.

		target (f.+m.)			target (f.)			percep.ass. (f.+m.)			percep.ass. (f.)		
		B	**M**	**P**	**B**	**M**	**P**	**B**	**M**	**P**	**B**	**M**	**P**
		Reference Features:											
MFCCs	B	**71**	18	11	**74**	17	9	**72**	23	5	**79**	19	2
	M	22	**56**	22	18	**63**	19	16	**64**	20	12	**73**	15
	P	20	33	**47**	12	42	**46**	11	46	**43**	7	54	**38**
VQ	B	**60**	29	11	**54**	25	21	**42**	37	21	**49**	27	24
	M	27	**47**	26	25	**42**	33	13	**60**	27	17	**52**	31
	P	23	45	**32**	17	35	**48**	9	41	**50**	11	31	**58**
ZFF	B	**63**	26	11	**64**	24	12	**41**	48	11	**47**	41	12
	M	28	**52**	20	30	**44**	26	19	**65**	16	21	**60**	19
	P	24	34	**42**	12	28	**60**	13	37	**50**	4	41	**55**
		Proposed Features:											
MPS_{peaks}	B	**79**	9	12	**90**	5	5	**80**	15	5	**85**	10	5
	M	6	**77**	17	4	**78**	18	9	**75**	16	8	**78**	14
	P	12	34	**54**	7	40	**53**	11	55	**34**	7	54	**39**
MPS_{sum}	B	**77**	12	11	**88**	5	7	**72**	23	5	**82**	13	5
	M	8	**76**	16	6	**74**	20	8	**76**	16	10	**74**	16
	P	12	32	**56**	9	27	**64**	7	46	**47**	6	47	**47**
$Ceps_{peaks}$	B	**76**	8	16	**82**	8	10	**79**	17	4	**87**	9	4
	M	8	**73**	19	10	**72**	18	8	**80**	12	8	**71**	21
	P	17	31	**52**	11	40	**49**	9	74	**17**	6	65	**29**

better or nearly similarly to FSC9, for both target and perceptually assessed labels, even though FSC9 holds some or all reference features. Note that the feature set combinations for each corresponding set of labels (target and perceptually assessed) include the best performing feature sets of the reference features and the proposed features. Thus, the feature set combinations vary for the different set of labels.

VII. DISCUSSION

From the results in Tables 4, 7, 10, and 11, it is clearly evident that the performance for the MPS_{sum} (8-dimensional) features is similar or better than the MPS_{peaks} (40-dimensional) features.

This is explainable by the summation included in calculating the MPS_{sum} features. The summed modulation power spectrum is calculated by summing all the energy along the temporal axis for each spectral modulation frequency τ, whereas the MPS_{peaks} features only hold information on the peak amplitudes within the search regions. The extracted modulation information contained in the MPS_{peaks} and MPS_{sum} features seems to increase the capability of distinguishing modal from the other phonation modes. The peaks extracted from the cepstrum, averaged over consecutive time frames, contain similar information as the peaks extracted from the summed MPS, but without the temporal modulation information (i.e., vocal vibrato), which decreases the performance compared to the MPS_{sum} and MPS_{peaks} features.

The confusion matrices for all the features (reference and proposed) with the target and perceptually assessed labels in the whole dataset and in the female-only dataset scenario are given in Table 13. These confusion matrices show that the proposed MPS feature sets and the MFCCs exhibit the best performance in classifying breathy phonation, while the MPS_{peaks} feature set performs best over the whole dataset for both the target and the perceptually assessed labels.

The pronounced difference between breathy and other phonation modes can be visualized in the averaged modulation power spectra (averaged over all singers for each corresponding phonation mode), depicted in Fig. 6.[3] Breathy phonation shows the lowest values for the temporal

[3]We subtracted a fitted first order polynomial to detrend the MPS data for better visualization.

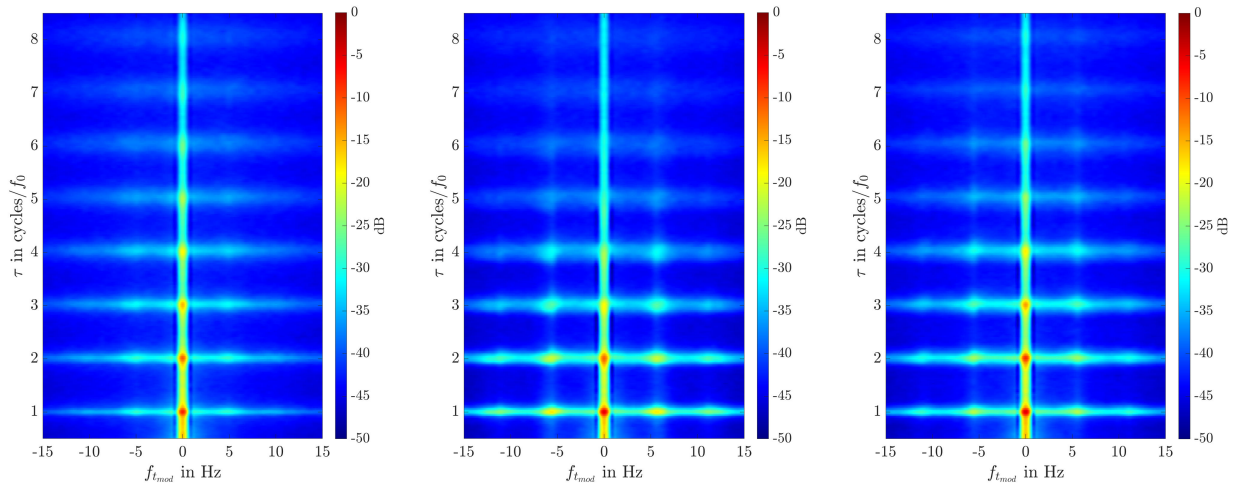


FIGURE 6. Averaged modulation power spectra of all singers (whole dataset) for breathy, modal, and pressed phonation. The modulation power spectrum for each sample is detrended by subtracting a fitted a first-order polynomial before averaging in order to highlight the differences between the phonation modes.

modulation components. Regarding the classification of pressed phonation, the confusion matrices of Table 13 overwhelmingly show a reduced performance for all investigated feature sets. The less distinct difference between modal and pressed phonation is also visible in the averaged MPS illustrated in Fig. 6. Nonetheless, slight differences between modal and pressed phonation modes are still detectable, especially at higher spectral modulation frequencies starting at $\tau = 3$ cycles/f_0, but they are less pronounced than in the visualizations presented in Fig. 3.

Additionally, Fig. 7[3] shows the summed MPS averaged over all singers for each corresponding phonation mode. Again, the discussed reduced differences between modal and pressed are visible. A decreased amplitude at $\tau = 1$ cycles/f_0, as seen in Fig. 4 for pressed phonation, diminishes in the averaged data.

The classification experiments with the perceptually assessed labels show a slightly lower performance compared to target labels for almost all feature sets. However, an approach other than using the median rating in combination with k-mediods clustering could lead to better performance, offering potential for future research. Overall, the results of the classification experiments show a higher mean accuracy for the novel MPS features when using the target labels. This suggests that the target labels, in combination with the new features, provide a better discrimination of phonation modes for the present dataset.

The variance of the classification accuracy is strongly influenced by the underlying data and the labels. The lowest standard deviations for all feature sets are present when the data is reduced to the female-only data and by using the labels from the perceptual listening assessment (see Table 11). This reduced variance is generally noticeable for the perceptually assessed labels, most likely due to the involved strategy, which entails categorizing these labels with the median value of six ratings. The variance seen for the target labels may

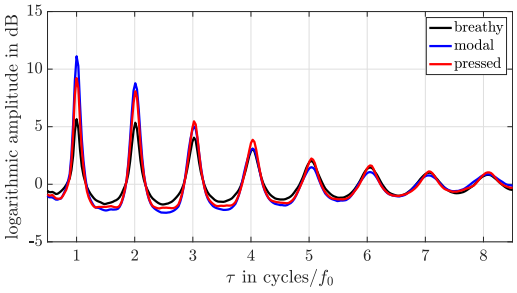


FIGURE 7. Averaged summed modulation power spectra of all singers for each phonation mode (breathy, normal, and pressed). The summed modulation power spectrum is detrended for each sample by subtracting a fitted first order polynomial for better visualisation.

be explained by the leave-one-singer-out strategy. A greater variance is to be anticipated if one or more singers perform the instructed phonation modes (target labels) with more consistency than the others. This is also evident in the experiments of the combined feature sets (for both target and perceptually assessed labels). Furthermore, the results demonstrate that there exists a weaker complementary information between the various feature sets.

In addition, a comparison between target labels and labels from the perceptual listening assessment (see Table 3) shows similar confusions (i.e., between modal and pressed phonations). This implies that either the singers were unable to fully reproduce the target phonation modes in the recordings or it was too challenging for listeners to distinguish the phonation modes. Most likely, both of these factors are at play. This limits, to some extent, the discussion on the performance of the current feature sets for classifying pressed phonation, and the comparison of performance between target or perceptually assessed labels. However, to the authors' knowledge, the data investigated in this work is currently the largest publicly available annotated dataset for phonation modes in singing and further investigations on the dataset should follow.

VIII. CONCLUSION

In this article, we have proposed three new feature sets based on the modulation power spectrum and the averaged cepstrum for the classification of phonation modes (breathy, modal, pressed) in classical singing. We have also presented a newly collected phonation modes dataset, which consists of ten classical singers singing several vowels at several pitches, which is publicly available at: https://phaidra.kug.ac.at/o:126552. Experiments were carried out on the whole (combination of female and male data) and the female-only data using target labels (instructed phonation modes during the recordings) as well as perceptually assessed labels (derived from a perceptual listening assessment). We have compared the proposed three features (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) with state-of-the-art reference features (VQ, ZFF, and MFCCs) in a phonation mode classification task. The results of the classification experiments reveal that the proposed features based on the MPS have a slightly better ability to accurately assess the phonation modes. In terms of performance and most striking in number of features, MPS_{sum} is the best performing feature set compared to other feature sets, which by itself produces a similar classification accuracy as the best performing feature set combination. Additionally, it has been found that the target labels result in a better performance than using the labels derived from the perceptually assessed ratings. It was found that the influence of the underlying data and the corresponding labelling play an important role in the classification. This should be taken into account in future approaches. Further research is still needed on deriving phonation mode specific features without using the fundamental frequency, as well as on thoroughly examining the listening assessment data.

IX. ETHICS AND CONSENT

This work involved human subjects or animals in its research. Ethical and experimental procedures and protocols were designed to comply with the proposal reviewed by the ethics advisory board of the University of Music and Performing Arts, Graz, and performed in line with the Helsinki declaration. All participants were informed that their participation was voluntary and could be withdrawn any time. Participants received an expense allowance for their voluntary participation.

REFERENCES

[1] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed—Automatic detection of phonation mode from audio recordings of singing," *J. New Music Res.*, vol. 42, no. 2, pp. 171–186, Jun. 2013, doi: 10.1080/09298215.2013.821496.

[2] J. Sundberg, *The Science of the Singing Voice*. DeKalb, IL, USA: Northern Illinois Univ. Press, 1987.

[3] H.-L. Lu and J. O. Smith, "Glottal source modeling for singing voice synthesis," in *Proc. ICMC*, 2000, pp. 1–8. [Online]. Available: http://hdl.handle.net/2027/spo.bbp2372.2000.186

[4] C. Gobl, "A preliminary study of acoustic voice quality correlates," *Quart. Prog. Status Rep.*, vol. 4, pp. 9–22, Jan. 1989.

[5] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Amer.*, vol. 31, no. 6, pp. 667–677, Jun. 1959, doi: 10.1121/1.1907771.

[6] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1625–1629, 1974, doi: 10.1121/1.1903487.

[7] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, pp. 109–118, Jan. 1992, doi: 10.1016/0167-6393(92)90005-R.

[8] P. Alku and E. Vilkman, "Preliminary experiences in using automatic inverse filtering of acoustical signals for the voice source analysis," *Scandin. J. Logopedics Phoniatrics*, vol. 17, no. 2, pp. 128–135, Jan. 1992, doi: 10.3109/14015439209098723.

[9] P. Alku, "Glottal inverse filtering analysis of human voice production—A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, Oct. 2011, doi: 10.1007/S12046-011-0041-5.

[10] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1117–1138, Sep. 2014, doi: 10.1016/j.csl.2014.03.003.

[11] I. Arroabarren and A. Carlosena, "Inverse filtering in singing voice: A critical analysis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1422–1431, Jul. 2006, doi: 10.1109/TSA.2005.858013.

[12] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 6, pp. 1170–1179, Jun. 2013, doi: 10.1109/TASL.2013.2245653.

[13] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Commun.*, vol. 118, pp. 33–47, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639319303358

[14] L. Rabiner and R. Schafer, *The Cepstrum and Homomorphic Speech Processing*. London, U.K.: Pearson, 2011.

[15] B. Bogert, M. Healy, and J. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and Saphe cracking," in *Proc. Symp. Time Ser. Anal.*, vol. 15, 1963, pp. 209–243.

[16] J. M. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech Hearing Res.*, vol. 37, no. 4, pp. 769–778, 1994, doi: 10.1044/JSHR.3704.769.

[17] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomed. Signal Process. Control*, vol. 14, pp. 42–54, Nov. 2014, doi: 10.1016/j.bspc.2014.07.001.

[18] D. Stoller and S. Dixon, "Analysis and classification of phonation modes in singing," in *Proc. 17th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, New York, NY, USA, 2016, pp. 80–86.

[19] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[20] J.-L. Rouas and L. Ioannidis, "Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings," in *Proc. Interspeech*, Sep. 2016, pp. 150–154, doi: 10.21437/Interspeech.2016-1135.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[22] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 7, pp. 1–8, Dec. 2003. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1155/S1110865703305013

[23] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[24] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 1647–1650.

[25] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 2514–2517, doi: 10.1109/IEMBS.2009.5334850.

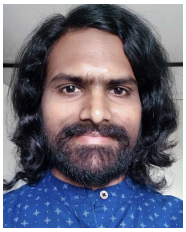
[26] M. Markaki and Y. Stylianou, "Modulation spectral features for objective voice quality assessment," in *Proc. 4th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, Mar. 2010, pp. 1–4, doi: 10.1109/ISCCSP.2010.5463313.

[27] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 1938–1948, Sep. 2011, doi: 10.1109/TASL.2010.2104141.

[28] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Amer.*, vol. 114, no. 6, pp. 3394–3411, 2003, doi: 10.1121/1.1624067.
[29] A. Hsu, S. M. N. Woolley, T. Fremouw, and F. E. Theunissen, "Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons," *J. Neurosci.*, vol. 24, pp. 9201–9211, Jan. 2004, doi: 10.1523/JNEUROSCI.2449-04.2004.
[30] R. Husson and C. E. Seashoee, "Psychology of the vibrato in voice and instrument," *Revue Musicologie*, vol. 19, no. 66, p. 115, May 1938, doi: 10.2307/925282.
[31] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neurosci. Biobehav. Rev.*, vol. 81, pp. 181–187, Oct. 2017, doi: 10.1016/j.neubiorev.2017.02.011.
[32] Z. Zhang, "Mechanics of human voice production and control," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. 2614–2635, Oct. 2016.
[33] M. Frank, D. Rudrich, and M. Brandner, "Augmented practice-room—Augmented acoustics in music education," in *Proc. 46th Conf. Fortschritte Akustik, Deutsche Gesellschaft Akustik*, vol. 46, Mar. 2020, pp. 151–154. [Online]. Available: https://www.dega-akustik.de/publikationen/online-proceedings/
[34] N. Klanjscek, L. David, and M. Frank, "Evaluation of an E-learning tool for augmented acoustics in music education," *Music Sci.*, vol. 4, Aug. 2021, Art. no. 20592043211037511, doi: 10.1177/20592043211037511.
[35] N. Meyer-Kahlen, D. Rudrich, M. Brandner, S. Wirler, S. Windtner, and M. Frank, "DIY modifications for acoustically transparent headphones," *J. Audio Eng. Soc.*, vol. 148, pp. 1–5, May 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20841
[36] L. L. Beranek, "Concert Hall acoustics—1992," *J. Acoust. Soc. Amer.*, vol. 92, no. 1, pp. 1–39, 1992.
[37] M. Skålevik, "Reverberation time—The mother of all room acoustic parameters," in *Proc. 20th Int. Congr. Acoustic, Integr. Comput.-Aided Eng.*, vol. 10, 2010, pp. 2508–2513.
[38] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, Feb. 2002, doi: 10.1121/1.1490365.
[39] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. Interspeech*, Aug. 2007, pp. 1410–1413.
[40] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter—A new method for quantification of the glottal flow," *Speech Commun.*, vol. 22, no. 1, pp. 67–79, 1997, doi: 10.1016/S0167-6393(97)00020-4.
[41] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ZTWCCs)," in *Proc. Interspeech*, Sep. 2018, pp. 232–236, doi: 10.21437/Interspeech.2018-2498.
[42] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *Proc. Interspeech*, Sep. 2018, pp. 441–445, doi: 10.21437/Interspeech.2018-2502.
[43] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. Interspeech*, Aug. 2013, pp. 1916–1920.
[44] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. Interspeech*, Sep. 2015, pp. 1324–1328.
[45] S. J. Young, D. K. J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[46] H. K. Kathania, S. Shahnawazuddin, W. Ahmad, and N. Adiga, "Role of linear, Mel and inverse-Mel filterbanks in automatic recognition of speech from high-pitched speakers," *Circuits, Syst., Signal Process.*, vol. 38, no. 10, pp. 4667–4682, Oct. 2019, doi: 10.1007/s00034-019-01072-7.
[47] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. ISMIR*, 2000. Accessed: Mar. 23, 2023. [Online]. Available: https://ismir2000.ismir.net/indexnoframes.html and https://ismir2000.ismir.net/papers/logan_paper.pdf
[48] D. Grzywczak and G. Gwardys, "Audio features in music information retrieval," in *Active Media Technology*, D. Ślęzak, G. Schaefer, S. T. Vuong, and Y.-S. Kim, Eds. Cham, Switzerland: Springer, 2014, pp. 187–199.
[49] A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

MANUEL BRANDNER received the M.S. degree in electrical engineering and audio engineering from the Graz University of Technology, in 2014. He is currently pursuing the Ph.D. degree with the Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria. He was a Scientific Project Assistant with the University of Music and Performing Arts Graz, in 2014, where he worked in the research fields of active noise control, acoustic measurement design, sound zoning, and machine learning. He was an university assistant, from 2018 to 2020, and has been a lecturer, since 2020. His research interests include singing voice analysis, musical acoustics, microphone array processing, and directivity analysis.

PAUL ARMIN BEREUTER received the dual master's degree in electrical engineering and audio engineering from the University of Music and Performing Arts Graz, and the Graz University of Technology, in 2022, and currently pursues his Ph.D. degree with the Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, where he is also employed as a Research Assistant. During his master's degree, he was a Student Assistant with the Signal Processing and Speech Communication Laboratory, Graz. His research interests include audio signal processing, music information retrieval, machine learning, and acoustics.



SUDARSANA REDDY KADIRI (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, and the Ph.D. degree from the Department of Electronics and Communication Engineering (ECE), International Institute of Information Technology Hyderabad (IIIT-H), Hyderabad, in 2018. He was a Teaching Assistant for several courses at IIIT-H, from 2012 to 2018. Since 2019, he has been involved in teaching and mentoring activities at Aalto University, Espoo, Finland. He was a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, from 2019 to 2021, where he is currently a Research Fellow. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience. He has published more than 60 research papers in peer-reviewed journals and conferences in his research areas. He is a reviewer of several reputed journals and conferences. He was awarded the Tata Consultancy Services (TCS) Fellowship for his Ph.D. degree.

ALOIS SONTACCHI received the Diploma degree in electrical engineering and audio engineering and the Ph.D. degree in technical science from the Graz University of Technology, in 1999 and 2003, respectively. In 1999, he joined the Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz (KUG). He directed the IEM, from 2010 to 2016. Since 2016, he has been a Professor of acoustics and audio engineering with KUG. His research interests include perceptual evaluation, music information retrieval, and spatial audio signal processing. He is a member of the Audio Engineering Society and the German Acoustical Society and the Scientific Board Member of DAFx and the Ambisonics Symposium.

4.2 Design of a Vowel and Voice Quality Indication Tool Based on Synthesized Vocal Signals

This work was published as:

P. A. Bereuter, F. Kraxberger, **M. Brandner** and A. Sontacchi. Design of a Vowel and Voice Quality Indication Tool Based on Synthesized Vocal Signals. *150th AES Convention, Audio Engineering Society Convention e-Brief 642*, 2021.

The idea and concept of this article were outlined by me, the first author and the second author. The first author and second author wrote the original draft of the manuscript with periodical contributions from me. The revision has been carried out by me and the fourth author, and the editing was done by the first and second author. Most of the programming was done by the first and second author with periodical contributions by me.

Audio Engineering Society

Convention e-Brief 642

Presented at the 150th Convention
2021 May 25–28, Online

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

Design of a Vowel and Voice Quality Indication Tool Based on Synthesized Vocal Signals

Paul A. Bereuter[1,2], Florian Kraxberger[1,2], Manuel Brandner[1,3], and Alois Sontacchi[1,3]

[1] *University of Music and Performing Arts, Graz, Austria*
[2] *Graz University of Technology, Austria*
[3] *Institute of Electronic Music and Acoustics, Graz, Austria*

Correspondence should be addressed to Paul A. Bereuter (paul.bereuter@gmail.com)

ABSTRACT

Voice disorders due to strenuous usage of unhealthy voice qualities are a common problem in professional singing. In order to minimize the risk of these voice disorders, vital feedback can be given by making aware of one's sung voice quality. This work presents the design task of a vowel and voice quality indication tool which can enable such a feedback. The tool is implemented in form of a VST plug-in. The plugin's interface provides a graphical representation of voice quality and vowel intelligibility by means of two 2D voice maps. The voice maps allow a graphical distinction of three voice qualities (modal, breathy or creaky), and the representation of a sung vowel within the formant space spanned by the first and second formant frequency. The design process includes (i) building a ground truth dataset by using a modified speech synthesizer, (ii) linear prediction analysis, and (iii) the visualisation of the estimated vowel and voice quality by means of the 2D voice maps. The plugin's code is available as open source to enable further development.

1 Introduction

While vowels, formed by the human vocal tract are a familiar concept, the voice quality is determined by the vocal folds' mode of operation, as defined in [1]. In a signal processing context this means that, by estimating the vocal tract (VT) filter based on vocal signals, information on present vowels can be obtained. Looking at the voice quality, the excitation signal present at the vocal folds needs to be considered. As proposed by [2], the excitation signal can be estimated by inverse application of a VT filter estimation on the vocal signal. A renowned method for VT filter estimation in speech signal processing is the *linear prediction analysis* (LPA). Based on LPA, a feedback tool allowing the indication of vowels and voice-qualities is proposed. To do so, feature spaces and class-boundaries are to be considered, which are typical aspects of a classification task. The basis of such a task is a profound dataset, which in this case is created using a vocal signal synthesis algorithm. The synthesis algorithm considers the voice qualities *modal*, *breathy* and *creaky* as well as five vowels */a/*, */e/*, */i/*, */o/* and */u/* and the fundamental frequency f_0 as input variables, allowing the creation of a labeled dataset.

The synthesis and analysis part of this work are discussed in section 2. Section 3 features details on the VST plugin implementation. The plugin's source code is publicly available on https://git.iem.at/PAB/lpa_voice_qual_eval.

2 Methods

2.1 Sung Vocal Signal Synthesis

A common model for the synthesis of human vocal signals is the *source-filter model*, where the source signal is responsible for the voice quality and the filter is assumed to be the vocal tract, responsible for the vowel. A model of the *source* or excitation signal $e[n]$ is given by the Liljencrants-Fant model (LF), which allows the modeling of source signals based on four parameters [3]. $e[n]$, designed by means of the LF-model with a sampling frequency of $f_s = 48\,\text{kHz}$, is basically a formulation of the derivative airflow through the glottis, also called derivative glottal flow (dGF). The *filter* is described by the transfer function of the vocal tract $H_{\text{VT}}(z)$, and defines the perceived vowel. It is modeled as an all-pole filter, with the transfer function

$$H_{\text{VT}}(z) = \frac{G}{\sum_{k=0}^{p} a_k z^{-k}} = \frac{S(z)}{V(z)}, \tag{1}$$

as defined in [4], where G is the VT gain, a_k are the filter coefficients, and $S(z)$ and $V(z)$, respectively, are the numerator and denominator polynomials.

An implementation of the LF-model in conjunction with an all-pole model of the vocal tract is given with in repository 1 of OPENGLOT [2]. The code of this repository was modified and extended to allow the creation of different signal realizations of equal voice quality and to better come up for the distinctive aspects of sung vocal signals in comparison to speech. In order to enable the LF-model parameter variations, the voice quality dependent distributions given by [1] are used to create different signal realizations with the same voice quality. Additionally, aspiration noise for breathy voice quality is added as suggested in [5]. Furthermore, in order to aim towards the aspects of sung vocal signals, vibrato is considered, i.e. jitter (frequency modulation) and shimmer (amplitude modulation) are applied onto the excitation signal according to [6]. Finally, the convolution of the excitation signal $e[n]$ with the impulse response of the vocal tract filter $H_{\text{VT}}(z)$ forms the synthesized vocal signal $s[n]$.

2.2 Parameter Variations

To evaluate class boundaries within the voice quality feature space mentioned in subsection 2.5, a labeled dataset is necessary, which provides a Monte-Carlo-style representation of the LF-model's parameter distributions from [1], 100 realizations of 0.5 s duration for each combination of the following parameters have been created: 5 vowels (*/a/*, */e/*, */i/*, */o/*, */u/*), 3 voice qualities (modal, breathy, creaky), and 10 fundamental frequencies ($f_0 \in \{70,\ 120,\ 170,\ \ldots,\ 520\}\,\text{Hz}$). This results in a dataset containing $5 \cdot 3 \cdot 10 \cdot 100 = 15000$ synthesized sung vocal signals which form the *measurement* dataset in subsection 2.5, wheras the *ground truth* dataset consists of 500 synthesized dGF-signals per voice quality and fundamental frequency.

2.3 Linear Prediction Analysis

The main goal of the algorithm's analysis part is to obtain relevant information on the vowel and voice quality. An estimation of the all-pole VT filter, delivering the information on the vowel, can be obtained by LPA. By applying the estimated all-pole VT filter inversely onto the sung vocal signal, the source-filter convolution is reverted and an estimation of the excitation signal $\hat{e}[n]$, holding information on the voice quality, is obtained. The signal flow and processing steps necessary to estimate the vocal tract filter, and further indicate the vowel and voice quality, are shown in Fig. 1.

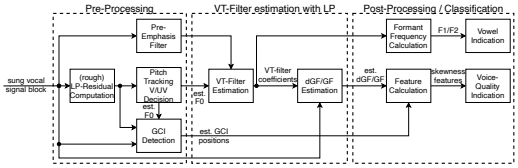


Fig. 1: General signal flow of vowel and voice quality indication tool (blockwise processing)

In order to ensure real-time processing, the signal flow depicted in Fig. 1 is processed in a blockwise manner with $t_{\text{block}} = 80\,\text{ms}$ block length and an overlap of $\text{OL} = 70\,\%$. The signals are downsampled by a factor of 3 before passing through the processing steps shown in Fig. 1, which leads to a downsampled sampling frequency of $f_s = 16\,\text{kHz}$ for the synthesized signals in the dataset described in subsection 2.2.

Pre-Processing. The first step is the computation of the residual signal $\hat{e}_r[n]$, which already includes the inverse filtering with a roughly estimated VT filter. The signal $\hat{e}_r[n]$ is used for the estimation of the fundamental frequency $\hat{f}_0$ and the detection of the glottal closure

instants, according to [7, 8]. Furthermore, the procedure in [7] allows to determine voiced and unvoiced signal blocks. The fundamental frequency $\hat{f}_0$ is then used in the VT filter estimation and the GCI estimates $\hat{n}_e$ are used in the classification features' calculation.

The last pre-processing step is the pre-emphasis filtering, whose goal it is to "whiten" the sung vocal signal. The whitening regularizes the linear prediction process, as the LPA was designed with preconceived white noise signals [9, p.4-8]. A simple first order IIR highpass-filter is chosen as the pre-emphasis filter.

Vocal Tract Filter Estimation. In order to obtain an accurate estimation of the VT filter, the *autocorrelation method using cepstral refinement* as proposed in [10] is used. This method can be seen as an improved variant of the classic autocorrelation method and is especially applicable for high-pitched voices. The signal model $\hat{s}[n] = \sum_{k=1}^{p} \hat{a}_k s[n-k]$, being a linear combination of coefficients $\hat{a}_k$ and past signal samples $s[n-k]$, in conjunction with the true signal $s[n]$, can be used to set up a Minimum Mean Square Error (MMSE) problem with the cost function J_{MSE}, in order to estimate the filter coefficients of the vocal tract filter [9, p.9-10]. A solution to this MMSE problem is given by the *Yule-Walker* equation system which can be solved by

$$\hat{\boldsymbol{a}}_{\mathrm{opt}} = \boldsymbol{R}_{ss}^{-1} \boldsymbol{r}_{ss+1}, \tag{2}$$

where $\boldsymbol{R}_{ss}$ is the symmetric autocorrelation matrix with Toeplitz structure, $\boldsymbol{r}_{ss+1}$ is an autocorrelation vector with a lag of one and $\hat{\boldsymbol{a}}_{\mathrm{opt}}$ are the estimated vocal tract filter coefficients. The autocorrelation function $r_{ss}[m]$ building $\boldsymbol{R}_{ss}$ and $\boldsymbol{r}_{ss+1}$, can be calculated in the frequency domain with

$$\begin{aligned} r_{ss}[m] &= \mathscr{F}^{-1}_{k\mapsto m}\left\{ \left| \mathscr{F}_{n\mapsto k}\{\tilde{s}[n]\cdot w[n]\}[k]\,/W \right|^2 \right\}[m] \\ W &= \textstyle\sum_{n=0}^{N_{\mathrm{block}}} w[n]/N_{\mathrm{block}} \\ N_{\mathrm{block}} &= \lfloor t_{\mathrm{block}} \cdot f_s \rceil\,, \end{aligned} \tag{3}$$

where W denotes a correction factor needed due to Hann-windowing with $w[n]$ of the pre-emphasis filtered signal block $\tilde{s}[n]$. $\mathscr{F}_{n\mapsto k}\{\star\}[k]$ denotes the discrete Fourier transform, and $\lfloor\star\rceil$ denotes rounding to the nearest integer.

As discussed in [9, p.11], an efficient algorithm to solve eq. 2, is the *Levinson-Durbin* algorithm. Solving the *Yule-Walker* equation system by means of the autocorrelation function in eq. 3 is also referred to as the *autocorrelation method* [9, p.11]. As shown in [10], limitations for the autocorrelation method arise with fundamental frequencies f_0 exceeding 200 Hz. For larger f_0, the dGF's spectral harmonic structure is present in the vicinity of the VT filter's poles, leading to misestimations with LPA. A way to counteract the excitation signal's spectral influence is cepstral smoothing, shown in [10].

This means, that the autocorrelation function used to build eq. 2, is transformed into the cepstral domain, where the excitation signal's portion, located at higher quefrencies, is liftered with an $\hat{f}_0$-dependent *Tukey*-window $w_{\mathrm{c}}[q]$. The refined autocorrelation function $\tilde{r}_{\mathrm{ss}}[m]$ is obtained by transforming the liftered cepstrum back into the time-domain as follows

$$\tilde{r}_{\mathrm{ss}}[m] = \mathscr{F}^{-1}_{k\mapsto m}\left\{ \left| \mathrm{e}^{\mathscr{F}_{q\mapsto k}\{c_{\mathrm{rss}}[q]\cdot w_{\mathrm{c}}[q]\}[k]} \right| \right\}[m]\,, \tag{4}$$

where $c_{\mathrm{rss}}[q]$ denotes the cepstrum of the autocorrelation function computed with

$$c_{\mathrm{rss}}[q] = \mathscr{F}^{-1}_{k\mapsto q}\{\ln|\mathscr{F}_{m\mapsto k}\{r_{\mathrm{ss}}[m]\}[k]|\}[q]\,. \tag{5}$$

The cepstrally refined autocorrelation $\tilde{r}_{\mathrm{ss}}[m]$ is then used to solve the *Yule-Walker* equation system with the *Levinson-Durbin* algorithm to obtain the estimated filter coefficients $\hat{\boldsymbol{a}}_{\mathrm{opt}}$. The filter gain is estimated as the square root of the remaining estimation error $J_{\mathrm{MSE}}\left(\hat{\boldsymbol{a}}_{\mathrm{opt}}\right)$, i.e. $\hat{G} = \sqrt{J_{\mathrm{MSE}}\left(\hat{\boldsymbol{a}}_{\mathrm{opt}}\right)}$.

The estimated filter coefficients $\hat{\boldsymbol{a}}_{\mathrm{opt}} = [1,\ \hat{a}_1,\ \hat{a}_2,\ \ldots,\ \hat{a}_p]$ and the filter gain $\hat{G}$ are inversely applied onto the original signal block, leading to an estimation of the dGF signal $\hat{e}[n]$, such that

$$\hat{e}[n] = \frac{1}{\hat{G}} \sum_{i=0}^{p} \hat{a}_i s[n-i] \quad \text{with} \quad \hat{a}_0 = 1. \tag{6}$$

2.4 Graphical Indication of Vowel

Sendlmeier *et al.* provide a vowel map based on the first two formants for long spoken German vowels [11] by means of mean and standard deviation values of the first two formants. A vowel map created based on their data, is visualized in the left subplot of Fig. 3. In order to indicate the vowel, estimated formant frequencies are plotted in the vowel map. From the estimated VT filter

coefficients $\hat{a}_i$, formant frequencies and bandwidths are evaluated, which are determined through the location of the corresponding pole in the complex plane. The formant frequency F_i and bandwidth B_i of the i-th pole $z_{p,i}$ are calculated as follows:

$$F_i = \tan^{-1}\left(\frac{\Im\{z_{p,i}\}}{\Re\{z_{p,i}\}}\right)\frac{f_s}{2\pi}, \quad B_i = \frac{-\ln|z_{p,i}|\, f_s}{\pi}. \quad (7)$$

F_i and B_i are checked against plausibility criteria, i.e. an upper bandwidth limit $B_{\max} = 500\,\text{Hz}$, a minimum formant frequency $F_{\min} = 90\,\text{Hz}$ and a maximum formant frequency $F_{\max} = 3.5\,\text{kHz}$. Formants which violate one of these criteria are omitted. The detected formants with the lowest two frequencies are considered to be the first two formants $\hat{F}_1$ and $\hat{F}_2$. The estimated formants of several signal blocks are plotted in the vowel map presented in Fig. 3, allowing for a graphical indication of the present vowel based on their location.

2.5 Graphical Indication of Voice Quality

In order to create a 2D voice quality map, a supervised learning problem is set up based on skewness-related measures of the estimated dGF signal $\hat{e}[n]$. Through numerical integration of $\hat{e}[n]$, the airflow through the glottis (*glottal flow*, GF) is obtained [3] and two skewness-related measures are evaluated: Firstly, the skewness of the dGF's amplitude values, and secondly, the skewness of the glottal flow. Together, they span a 2D feature space, allowing the distinction of different voice qualities.

Each signal of the synthesized dataset described in subsection 2.2 is analyzed, and the skewness measures are evaluated. A Support vector machine (SVM) is employed to provide class boundaries for the three voice qualities. To incorporate a possible fundamental frequency dependence of the analysis algorithm, the dataset is sub-divided as follows: The i-th data subset consists of data for the lowest fundamental frequency $f_0 = 70\,\text{Hz}$ and holds the estimated skewness values of all realizations up to the i-th fundamental frequency of $f_0 \in \{70,\ 120,\ 170,\ \ldots,\ 520\}\,\text{Hz}$. This is called the *measurement* dataset for the i-th fundamental frequency. Additionally, skewness measures of synthesized ground truth dGF-signals are evaluated, forming the *ground truth* dataset. Both the measurement and the ground truth dataset are randomly split into 80 % training data and 20 % test data. The SVM is trained with the training data of both *measurement* and *ground truth* datasets.

The training and test prediction score for each upper frequency limit $f_{0,i}$ is evaluated. In Fig. 2 it can be observed that the prediction scores are dependent on the fundamental frequency range. For fundamental frequencies $f_0 \in [70, 320]\,\text{Hz}$, prediction scores exceeding 90 % can be achieved. The SVM trained with data for this frequency range is used to classify points of a mesh grid sampling the voice quality feature space in order to visualize the class boundaries, leading to the 2D voice quality map displayed in the right subplot of Fig. 3.

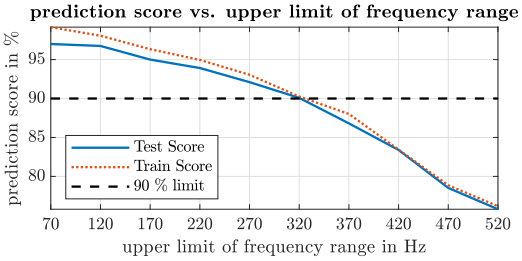


Fig. 2: SVM test and training score

3 VST Implementation

The autocorrelation method using cepstral refinement mentioned in subsection 2.3, and the created 2D vowel and voice quality maps discussed in section 2 are implemented in a VST plugin using C++ and the JUCE framework [12]. The plugin reads and buffers signals from a digital audio workstation and processes the steps depicted in Fig. 1. The features for vowel and voice quality indication are extracted for each signal block, and the results are plotted onto the 2D maps, which were previously created according to subsections 2.4 and 2.5. The results of 15 previous signal blocks are plotted onto the 2D maps as a trace of black dots as visible in Fig. 3. The results shown in Fig. 3 belong to a synthesized, modal */a/* vowel at $\hat{f}_0 \approx 300\,\text{Hz}$. The buffer structure allowing block processing was taken from [13]. An additional field in the GUI's lower right corner is indicating the current $\hat{f}_0$, calculated during the preprocessing steps of the analysis stage. The only variable parameter is the LPA-order (i.e., p in eq. 6) which determines how many coefficients are assumed for the VT filter estimation.

4 Summary

This work provides insight into the design process of a vowel and voice quality indication tool. Using a sung

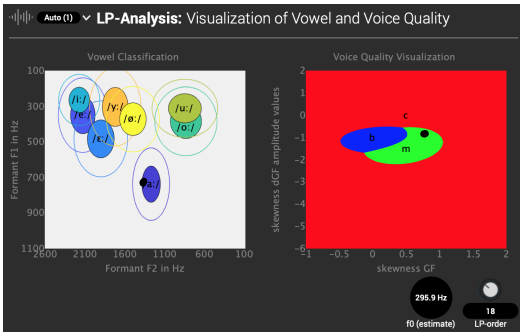


Fig. 3: GUI of proposed indication tool

vocal signal synthesizer, a dataset consisting of 15000 synthesized vocal signals and 15000 synthesized dGF-signals, with voice quality and vowel labels, are created. The *autocorrelation method using cepstral refinement* was chosen for the linear prediction analysis of the signal realizations, enabling the calculation of vowel and voice quality indication features. With the calculated voice quality features, and a SVM, a 2D voice quality feature map which visualizes the class boundaries for modal, breathy and creaky voice of the synthesized signals is created. The prediction score shown in Fig. 2 also indicates a dependence on the fundamental frequency range. In order to provide a graphical indication of the sung vowel, existing data from [11] was used to create a vowel map spanned by the first two formant frequencies. The created 2D maps and the discussed LPA algorithm were implemented into a VST plugin, allowing a real-time indication of the current sung vowel and voice quality. As the design process is carried out for synthesized signals, the created voice quality map is only useful for synthesized sung vocal samples. Nevertheless, with a comprehensive dataset consisting of vocal signals with labeled voice quality, sung by real singers, the proposed method provides a blueprint on how to implement a visual vowel and voice quality feedback tool, not only for synthesized vocal signals, but also for recordings of real singers.

References

[1] Gobl, C., "A Preliminary Study of Acoustic Voice Quality Correlates," in *STL-QPSR*, volume 30(4), pp. 9–22, KTH Stockholm, 1989.

[2] Alku, P., Murtola, T., Malinen, J., Kuortti, J., Story, B., Airaksinen, M., Salmi, M., Vilkman, E., and Geneid, A., "OPENGLOT - An open environment for the evaluation of glottal inverse filtering," *Speech Commun.*, 107, pp. 38 – 47, 2019.

[3] Fant, G., Liljencrants, J., and Lin, Q., "A Four-Parameter Model of Glottal Flow," in *STL-QPSR*, volume 26(4), pp. 1–13, KTH Stockholm, 1985.

[4] Gold, B. and Rabiner, L. R., "Analysis of digital and analog formant synthesizers," *IEEE Trans. Audio Electroacoust.*, 16(1), pp. 81–94, 1968.

[5] Lu, H.-L. and Smith, J. O., "Glottal source modeling for singing voice synthesis," 2000, CCRMA, Stanford University.

[6] Sciri, P., *Singing Voice Vibrato: Measurement and Modification*, Master's thesis, IEM, University of Music and Performing Arts, Graz, 2011.

[7] Drugman, T. and Alwan, A., "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," in *Proc. INTERSPEECH 2011*, pp. 1973–1976, Florence, Italy, 2011.

[8] Drugman, T., Thomas, M., Guðnason, J., Naylor, P., and Dutoit, T., "Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review," *IEEE Trans. Audio Speech Lang. Process.*, 20(3), pp. 994–1006, 2012.

[9] Kovacevic, B., Milosavljević, M., Veinović, M., and Markovic, M., *Robust Digital Processing of Speech Signals*, Springer, 2017.

[10] Rahman, M. S. and Shimamura, T., "Linear Prediction Using Refined Autocorrelation Function," *EURASIP J. Audio Speech*, 2007(1), p. 045962, 2007.

[11] Sendlmeier, W. F. and Seebode, J., "Formantkarten des deutschen Vokalsystems," *TU Berlin, Institut für Sprache und Kommunikation*, 2006.

[12] Storer, J. and ROLI, "JUCE," Online-Resource, 2021, version 6.0.3, [accessed 29.04.2021] , `https://juce.com/`.

[13] Holzmüller, F., Bereuter, P., Merz, P., Rudrich, D., and Sontacchi, A., "Computational Efficient Real-Time Capable Constant-Q Spectrum Analyzer," in *AES Convention e-Brief 567*, Audio Eng. Soc., 2020.

View publication stats

5 Concluding Remarks

In the areas of voice directivity analysis and phonation mode analysis, this doctoral dissertation investigated new objective parameters for the analysis of the classical singing voice. The measurement setup created enables in-depth analysis of a singer's voice directivity and the extraction of other data, such as the mouth opening (`https://opendata.iem.at/projects/dirpat/`). The measurement setup was examined in detail using a head and torso simulator. This made it possible to address differences in the measured data of singers more clearly. The information about the mouth opening enabled a comparison of the simulation and the measurement-based data, and examination of the effects of vowel- or singing-style-related changes in mouth opening on voice directivity. The findings demonstrated a clear relationship between voice directivity and mouth opening. The dataset created in the course of this work to investigate the classical singing voice directivity consisting of ten classical singers is made available publicly (`https://phaidra.kug.ac.at/o:127284`).

Audio samples with various voice directivity patterns for speech and noise sounds have been generated in a virtual environment to investigate how changes in voice directivity are perceived. These samples were utilized in a listening test. The results indicate a strong relationship of voice directivity on the direct-to-diffuse ratio, but also show that perceived sound is highly dependent on the listener's position and on the spectral composition of sound. The changes are perceived more prominently for noise than speech. Differences are less well perceived if the listener is not positioned towards the sound source and aligned on-axis. The effects of changes in voice directivity on the perception of one's own voice are less pronounced when the singer sings into the open space than when standing close to a reflective surface or wall. The two contributions in this thesis on perception of voice directivity in combination with the measurement results suggest that changes in voice directivity in classical singing play a subordinate role in auditory perception.

In order to categorize phonation modes in classical singing, new features and data were presented as a result of the complementary investigation of phonation modes. The newly developed features extracted from the modulation power spectrum perform better than reference features derived by existing techniques. These results provided insightful information about alternative signal processing approaches for classifying phonation modes. The contribution also addresses the process of collecting and labeling new data for classical singing analysis. The development of a VST-plugin as a software tool for vocal analysis also contributed to the discovery of current limitations in the parameter extraction from the estimated glottal waveform in classical singing, and to the demonstration of a possible visual layout for phonation mode classification and vowel identification.

In the areas of voice directivity and voice quality analysis, this thesis advances recent research on the analysis of the classical singing voice. Preliminary research on different signal processing techniques to extract information for singing voice analysis helped to

better understand the current limitations of the popularly available approaches. The current limitations in classical singing analysis are primarily due to the fact that most voice analysis techniques are built on speech-optimized algorithms, along with the lack of freely available, high-quality audio recordings for research and algorithm development. This thesis makes a contribution by outlining novel techniques for characterizing features specifically relevant to the analysis of the classical singing voice and by making the newly generated data from the cited publications accessible to the public.