Christoph Pörschmann, Johannes M. Arend

Analysis and visualization of dynamic human voice directivity

Analysis and Visualization of Dynamic Human Voice Directivity

Christoph Pörschmann[1], Johannes M. Arend[1,2]
[1] TH Köln, Institute of Communications Engineering, Cologne, Germany
[2] TU Berlin, Audio Communication Group, Berlin, Germany
Email: christoph.poerschmann@th-koeln.de

Introduction

In many everyday situations, we experience the influence of the human voice directivity. We perceive loudness and timbre differently when a speaker faces us or turns away from us. Often, we use voice directivity intuitively, for example when facing a person in a meeting or a casual conversation. Such effects of human voice directivity have long been a topic of research. Early studies were carried out more than 200 years ago analyzing the directional radiation of speech in general [1, 2, 3]. In 1929 a first approach determining directivity patterns for several vowels and fricatives in the horizontal plane has been presented by Trendelenburg [4]. In 1939 Dunn and Farnsworth [5] determined spherical directivity patterns for a spoken sentence at different distances in third-octave bands from 63 Hz up to 12 kHz. Since then, voice directivity has been subject to many studies, either applying human speakers or dummy heads with integrated mouth simulators.

A specific characteristic of the human voice that cannot be analyzed with a dummy head is its dynamic directivity, i.e., time-variant changes that occur when speaking or singing. To adequately determine these features, the sound radiation needs to be captured for an appropriately large number of directions. Katz and D'Alessandro [6] analyzed the voice directivity in the horizontal plane for sustained vowels articulated by a professional opera singer. Even though the study showed no systematic differences between the different vowels, it gave first insights into the vowel-dependencies. Kocon and Monson [7] examined articulation-dependent effects of voice directivity for different vowels in fluent speech. They observed an effect of the vowel on the directivity pattern and determined the strongest directivity for an [a]. Monson et al. [8] analyzed phoneme-dependencies in the horizontal plane for measurements of the directivity in angular steps of 15°. They showed that the directivity varies strongly for different articulations, e.g., between voiceless fricatives with a more directive sound radiation of an [s] than of an [f]. However, this study did not show significant differences between speech and singing, and only minor dependencies on the articulation level. In contrast, Chu and Warnock [9] observed significant differences depending on the articulation level. Postma and Katz [10] as well as Postma et al. [11] analyzed the influence of dynamic voice directivity for auralizations. Their results indicate that auralizations involving dynamic voice directivity are perceived as more plausible and exhibit a wider apparent source width than auralizations with static voice directivity or omnidirectional sources. In contrast, a recent study by Ehret et al. [12] concluded that dynamic directivity is perceptually indistinguishable from a static voice directivity.

In general, voice directivity measurements can either be performed sequentially for an arbitrary number of directions or simultaneously using a surrounding microphone array. In the case of sequential measurements, the spectrum of the radiated sound is typically analyzed averaged over time. Thus, time-variances influencing the speaker's directivity can hardly be resolved and are in most evaluations not considered. For determining the spherical directivity patterns, it is advantageous to apply surrounding microphone arrays (SMAs) [13, 14, 15, 16]. Generally, the setup is restricted to a limited number of sampling points. Thus, the spatial resolution of the directivity sets is sparse. Consequently, methods are required for spatial upsampling of the (sparsely) measured datasets by appropriate interpolation between the measured directions.

Up to now, only a limited number of studies on spherical voice directivity have been published (e.g., [17, 9, 15]), but none of them analyzed the time-variant effects of fluent speech. This pilot study aims to investigate to what extent high-density directivity sets can be determined from fluent speech measured on a sparse sampling grid with an SMA. In this context, we presented the SUpDEq (Spatial Upsampling by Directional Equalization) method [18], which originally was designed for spatial upsampling of head-related transfer functions. In Pörschmann and Arend [19, 20], we evaluated the method for a dummy head with mouth simulator and showed that reasonable dense directivity sets can be obtained from sparse measurements. Our studies revealed that measurements in a surrounding microphone array with 32 microphones are sufficient to generate a decent full-spherical dense directivity set, with an error averaged over the entire sphere below 4 dB for frequencies up to 8 kHz. In [21, 22], we applied the SUpDEq method to human voice directivity and determined full-spherical directivity patterns of five vowels and three fricatives for 13 persons using a surrounding spherical microphone array with 32 microphones. The results showed significant differences between the two groups of phonemes, and between some of the phonemes of each group. Furthermore, in this context we studied the influence of hand postures on voice directivity and showed that cupping the hands around the mouth or holding a hand in front of the mouth have stronger effects than differences between the phonemes [23]. In this paper, we present a pilot study that investigates dynamic voice directivity of fluent speech and thus resolves in which way voice directivity changes over time.

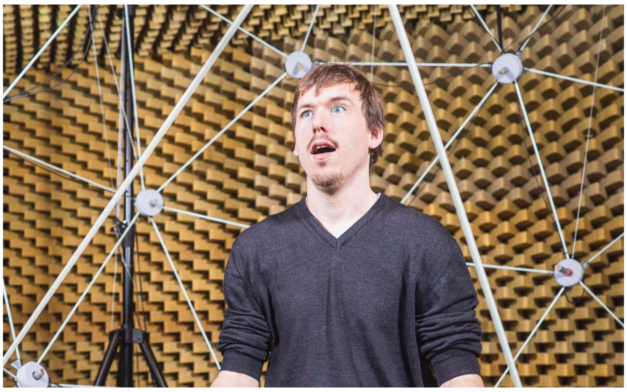

Figure 1: Human speaker inside the SMA during the measurements.

Method and Materials

Measurements

All measurements of this pilot study were performed in the anechoic chamber of TH Köln, sized 4.5 m × 11.7 m × 2.3 m (W×D×H) and with a lower cut-off frequency of about 200 Hz. The directivity patterns were measured with an SMA having a basic shape of a pentakis dodecahedron with 32 Rode NT5 cardioid microphones located at the vertices of this shape on a constant radius of 1 m. This sampling scheme allows resolving the directivity up to a spatial order of $N = 4$ [13]. In the present study, an additional Rode NT5 microphone was positioned at the front serving as a reference and for spectral equalization in postprocessing. Four RME Octamic II devices served as preamplifiers and AD / DA converters for the 32 microphones of the SMA. All signals of the SMA were processed with two RME Fireface UFX audio interfaces. For a more detailed description of the SMA setup, please refer to Arend et al. [14, 16]. In the present study, one of these audio interfaces was also used as a preamplifier and AD / DA converter for the reference microphone.

As test material, we recorded excerpts from Antoine de Saint-Exupery's book ”The little prince” [24], spoken by two German speakers in English and German. To obtain a visual representation of the time-variant changes of the mouth, we filmed the speakers using a Sony PXW-FS7 camera at a frame rate of 180 fps. The video can be superimposed with the determined directivity patterns or used to create slow-motion sequences of mouth movements. Fig. 1 shows the SMA with a person inside.

Postprocessing

The postprocessing is only briefly summarized here, as it was carried out similarly as presented in Pörschmann and Arend [20, 22]. To eliminate the influence of reflections and of room modes of the anechoic chamber, which in our case become prominent below 200 Hz, a low-frequency extension was applied substituting the original low-frequency component in the frequency domain by an adequately matched one of an analytic low-frequency model. Furthermore, in the postprocessing, the inaccuracies in positioning the subjects in the center of the microphone array were compensated. As small deviations of some centimeters already lead to strong impairments in the spatial upsampling process [25], we applied a method for distance error compensation that reduces the impairments of distance errors of the measured directivity patterns [20]. Then the compensated multichannel recordings were partitioned into frames of 67 ms (3200 samples at 48 kHz sampling rate) and Hann-windowed (50 % overlap). Further processing was done separately for each audio frame. The sparse datasets were spatially upsampled to a dense grid with 2702 sampling points on a Lebedev sampling scheme applying the SUpDEq method[1], which we described in detail in Pörschmann et al. [18] and evaluated for upsampling voice directivity in Pörschmann and Arend [20]. However, in contrast to the processing used in the studies mentioned above, we applied the postprocessing and the upsampling not to transfer functions but directly to the multichannel audio frames. As output of the processing chain, a high-density dataset was stored for each audio frame, which can be used for the further analysis of dynamic voice directivity.

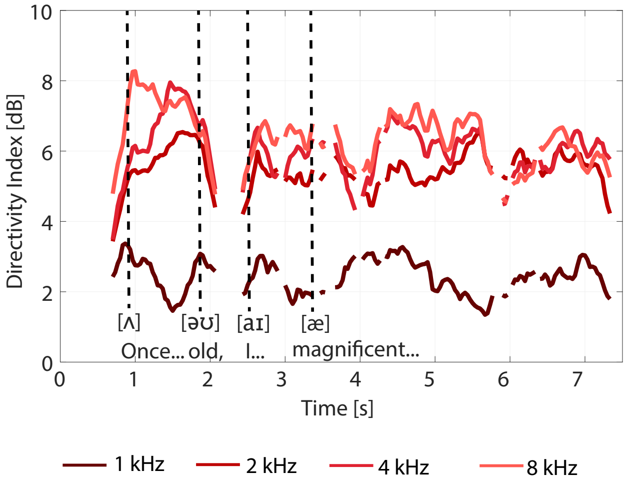

Figure 2: Directivity index (DI) over time for one sentence for octave bands of 1 kHz, 2 kHz, 4 kHz, and 8 kHz. Four selected phonemes are marked on the time axis.

Results

The results presented here are based on one selected sentence of the recordings, which was the first sentence of [24]: ”Once when I was six years old I saw a magnificent picture in a book, called True Stories from Nature, about the primeval forest.” In a first step, we calculated the directivity index (DI) for each of the frames:

$$DI(f,n) = 10\,lg\frac{4\pi|p(\phi_0,\theta_0,f,n)|^2}{\int\limits_0^{2\pi}\int\limits_{-\pi/2}^{\pi/2}|p(\phi,\theta,f,n)|^2 cos\theta d\theta d\phi}, \qquad (1)$$

with ϕ the azimuth, θ the elevation, ϕ_0, θ_0 the frontal direction, f the frequency, and n the index of the frame.

[1] A Matlab-based implementation of the SUpDEq method can be accessed on `https://github.com/AudioGroupCologne/SUpDEq`

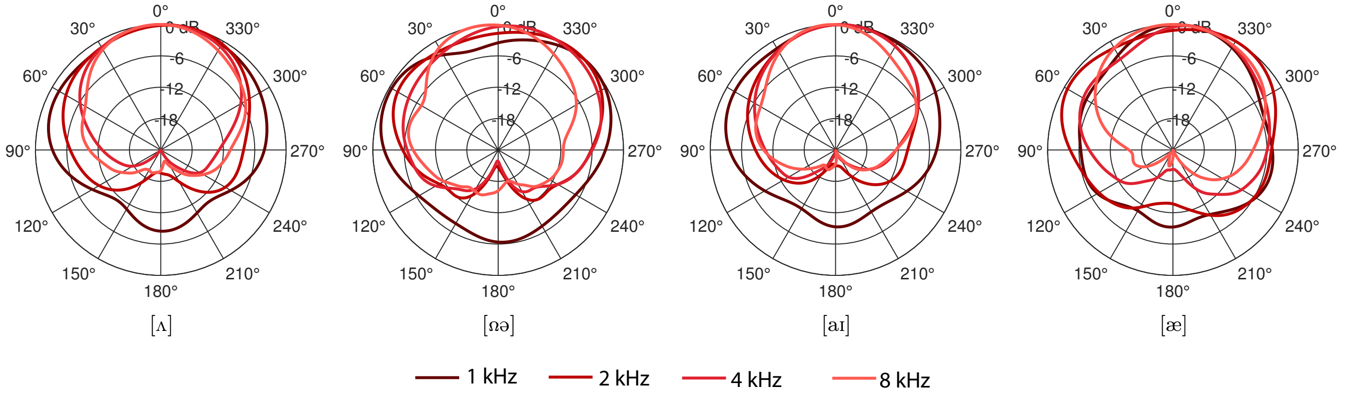

Figure 3: Directivity in the horizontal plane for an [ʌ], an [ʊə], an [aɪ], and an [æ] in octave bands of 1 kHz, 2 kHz, 4 kHz, and 8 kHz normalized to a maximal value of 0 dB.

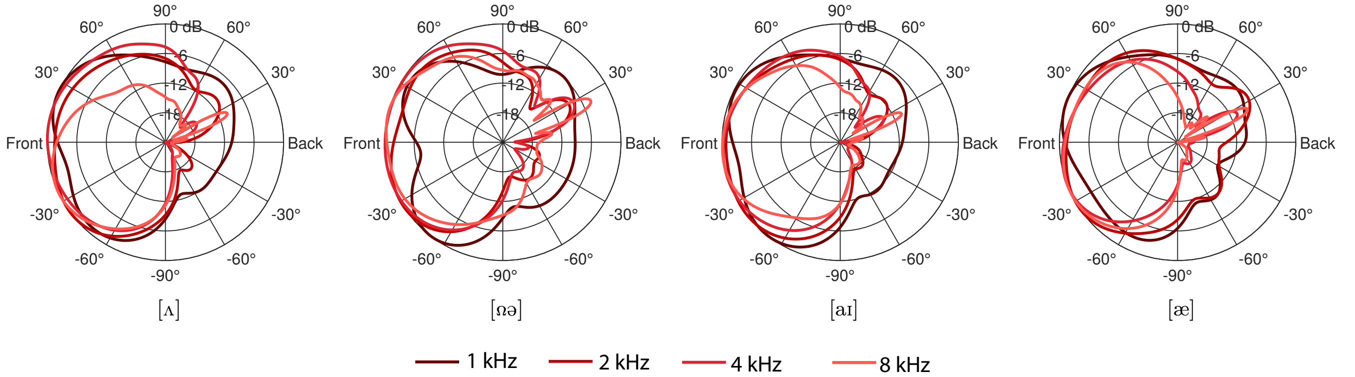

Figure 4: Directivity in the vertical plane for an [ʌ], an [ʊə], an [aɪ], and an [æ] in octave bands of 1 kHz, 2 kHz, 4 kHz, and 8 kHz normalized to a maximal value of 0 dB.

Fig. 2 shows the DI over time in octave bands. Since no stable values were obtained for speech pauses and sections with low energy, we refrained from plotting frames with an energy of 14 dB or more below the average energy of the entire sentence. Generally, the DI increases with frequency, which is in line with other studies [15, 22] and can easily be explained by frequency-dependent diffraction effects of the head. Furthermore, we found variations in DI over time caused by phoneme-dependencies. These variations occurred in a similar way as in our previous studies [21, 22], in which we analyzed directivity patterns of a variety of separately articulated phonemes. For our test sentence, we observed the highest DI for an [ʌ] at the beginning of the sentence.

Then we determined directivity patterns for selected phonemes within the sentence. Fig. 3 and 4 show the directivity patterns in the horizontal and vertical plane, for four different time frames in which an [ʌ], an [ʊə], an [aɪ], and an [æ] were spoken. The respective time frames are also marked in Fig. 2. It can be observed that significant differences between the phonemes occur in the horizontal plane. While for an [ʌ] and an [aɪ], the directivity patterns in the 4 kHz and 8 kHz octave band are very similar, they are stronger directed to the frontal direction for an [ʊə] at 8 kHz. Furthermore, the plots show slight asymmetries, e.g., for the [ʊə] or the [æ]. In the vertical plane, the frequency-dependent differences seem to be smaller than in the horizontal plane, at least in the frontal hemisphere. In the vertical plane, both, the variations between the phonemes as well as between the different frequency bands, tend to be reduced. However, in future work, these findings need to be examined in more detail based on larger parts of the test material.

Conclusion

In this pilot study, we analyzed dynamic voice directivity of fluent speech. We presented and demonstrated a method to determine high-density directivity sets from sparse measurements carried out with an SMA with 32 microphones. The results reveal how the dynamic directivity changes in fluent speech over time and how this affects the DI. Furthermore, superimposing the directivity plots with a filmed sequence of the subject allows studying the dynamics of the lip movements as well as of the form and the size of the mouth opening. The results presented in this pilot study are only based on one single sentence. Next, we plan to systematically segment and evaluate larger recordings. This will allow for a more detailed analysis of variations and fluctuations of the directivity. For this, processing tools for speech segmentation need to be applied and appropriately adapted.

The results of this study are of high relevance for different purposes. First, it is of general interest when studying voice production to analyze the dynamic voice directivity in detail for fluent speech. Second, methods and datasets are required for applications in the field of virtual reality, augmented reality, or room acoustic simulation, to integrate adequate voice radiation patterns into the process of sound-field calculation. The question of whether and to what extent time-variant aspects have to be considered has a strong influence on the system design and the required methods for obtaining the directivity datasets. Finally, when reproducing one's own voice in a virtual acoustic environment [26, 27, 28, 29], its dynamic directivity could be perceptible for the speaker himself.

Acknowledgements

We thank Raphael Gillioz and Kai Altwicker for their work on the measurements and the segmentation of the raw recordings. The research has been carried out in the research project NarDasS, funded by the Federal Ministry of Education and Research in Germany, support code: BMBF 03FH014IX5-NarDasS.

References

[1] Saunders, G., *Treatise on Theaters*, I. and J. Taylor, London, 1790.

[2] Wyatt, B., *Observation on the Design for the Theatre Royal, Drury Lane*, J. Taylor, London, 1813.

[3] Henry, J., "Annual Report of the Board of Regents of the Smithsonian Institution," Technical report, A. G. F. Nicholson, Washington, DC, 1857.

[4] Trendelenburg, F., "Beitrag zur Frage der Stimmrichtwirkung," *Zeitschrift für techn. Physik*, 11, pp. 558–563, 1929.

[5] Dunn, H. K. and Farnsworth, D. W., "Exploration of pressure field around the human head during speech," *The Journal of the Acoustical Society of America*, 10, pp. 184–199, 1939, doi: https://doi.org/10.1121/1.1915975.

[6] Katz, B. and D'Alessandro, C., "Directivity measurements of the singing voice," in *Proceedings of the 19th International Congress on Acoustics*, 2007.

[7] Kocon, P. and Monson, B. B., "Horizontal directivity patterns differ between vowels extracted from running speech," *The Journal of the Acoustical Society of America*, 144(1), pp. EL7–EL12, 2018, doi:10.1121/1.5044508.

[8] Monson, B. B., Hunter, E. J., and Story, B. H., "Horizontal directivity of low- and high-frequency energy in speech and singing," *The Journal of the Acoustical Society of America*, 132(1), pp. 433–441, 2012, doi:10.1121/1.4725963.

[9] Chu, W. T. and Warnock, A. C. C., "Detailed Directivity of Sound Fields Around Human Talkers," Technical report, 2002, doi:10.4224/20378930.

[10] Postma, B. N. J. and Katz, B. F., "Dynamic voice directivity in room acoustic auralizations," in *Proceedings of the 42th DAGA*, pp. 352–355, 2016.

[11] Postma, B. N. J., Demontis, H., and Katz, B. F. G., "Subjective Evaluation of Dynamic Voice Directivity for Auralizations," *Acta Acustica united with Acustica*, 103(2), pp. 181–184, 2017.

[12] Ehret, J., Stienen, J., Brozdowski, C., Bönsch, A., Mittelberg, I., Vorländer, M., and Kuhlen, T. W., "Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech," in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8, ACM, New York, NY, USA, 2020, doi: 10.1145/3383652.3423863.

[13] Pollow, M., *Directivity Patterns for Room Acoustical Measurements and Simulations*, Logos Verlag Berlin, 2015.

[14] Arend, J. M., Stade, P., and Pörschmann, C., "Binaural reproduction of self-generated sound in virtual acoustic environments," in *Proceedings of the 173rd Meeting of the Acoustical Society of America*, volume 30, pp. 1–13, 2017, doi: 10.1121/2.0000574.

[15] Brandner, M., Frank, M., and Rudrich, D., "DirPat - Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers," in *Proceedings of the 144th AES Convention, e-Brief 425*, 1, pp. 1–5, 2018.

[16] Arend, J. M., Lübeck, T., and Pörschmann, C., "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," in *Proceedings of the AES Conference on Immersive and Interactive Audio*, 2019.

[17] Kob, M. and Jers, H., "Directivity measurement of a singer," *The Journal of the Acoustical Society of America*, 105, p. 1003, 1999, doi:10.1121/1.425813.

[18] Pörschmann, C., Arend, J. M., and Brinkmann, F., "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(6), pp. 1060 – 1071, 2019, doi:10.1109/TASLP.2019.2908057.

[19] Pörschmann, C. and Arend, J. M., "A Method for Spatial Upsampling of Directivity Patterns of Human Speakers by Directional Equalization," in *Proceedings of the 45th DAGA*, pp. 1458 – 1461, 2019.

[20] Pörschmann, C. and Arend, J. M., "A Method for Spatial Upsampling of Voice Directivity by Directional Equalization," *Journal of the Audio Engineering Society*, 68(9), pp. 649–663, 2020, doi:10.17743/jaes.2020.0033.

[21] Pörschmann, C. and Arend, J. M., "Analyzing the Directivity Patterns of Human Speakers," in *Proceedings of the 46th DAGA*, pp. 1141 – 1144, 2020.

[22] Pörschmann, C. and Arend, J. M., "Investigating phoneme-dependencies of spherical voice directivity patterns," *The Journal of the Acoustical Society of America*, 149(6), pp. 4553 – 4564, 2021, doi:10.1121/10.0005401.

[23] Pörschmann, C. and Arend, J. M., "Effects of hand postures on voice directivity," *JASA Express Letters*, 2(3), p. 035203, 2022, doi:10.1121/10.0009748.

[24] de Saint-Exupery, A., *The Little Prince*, Reynal & Hitchcock, 1943.

[25] Pörschmann, C. and Arend, J. M., "How positioning inaccuracies influence the spatial upsampling of sparse head-related transfer function sets," in *Proceedings of the International Conference on Spatial Audio - ICSA 2019*, pp. 1–8, 2019.

[26] Pörschmann, C., "One's own voice in auditory virtual environments," *Acta Acustica united with Acustica*, 87(3), pp. 378–388, 2001.

[27] Pörschmann, C. and Pellegrini, R. S., "3-D Audio in Mobile Communication Devices: Effects of Self-Created and External Sounds on Presence in Auditory Virtual Environments," *JVRB - Journal of Virtual Reality and Broadcasting*, 7(11), pp. 3–11, 2010.

[28] Neidhardt, A., "Detection of a nearby wall in a virtual echolocation scenario based on measured and simulated OBRIRs," in *Proceedings of the AES Conference on Spatial Reproduction*, 2018.

[29] Frank, M. and Brandner, M., "Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions," in *Proceedings of the International Conference on Spatial Audio - ICSA 2019*, September, pp. 1–5, 2019.