forum acusticum 2026
8-12 September · Graz, Austria

PERCEPTUAL RELEVANCE OF SOURCE WIDTH IN BINAURAL RENDERING OF HUMAN SPEAKERS

Tobias Weber[1,2,*], Hendrik Himmelein[1,2,*], Christoph Pörschmann[1]

[1] *Institute of Computer and Communication Technology, TH Köln, Germany*
[2] *Audio Communication Group, TU Berlin, Germany*

Abstract

In recent years, numerous studies have investigated the plausibility of binaural rendering. Typically, the auralized sound sources are assumed to have point-source characteristics and static directivity. While this can be considered a good representation of loudspeakers, which typically have a well-defined acoustic center and time-invariant directivity, other natural sound sources, such as the human voice, may have different properties. This study presents the results of an initial listening experiment that determined the relevance of the 'source width' aspect when considering the sound radiation of human speakers. The experiment assessed static binaural recordings of frontally positioned human speakers in an anechoic environment using several attributes from the Spatial Audio Quality Inventory (SAQI). Two human speakers each articulated two different sentences at distances of 1.5 m and 3 m to the listener. The stimuli were compared to the reference of the same human speakers themselves articulating the same sentence. Visual cues were removed by placing a curtain between the speakers and the listeners. Two variants were analyzed: in the first, the binaural recording was presented dichotically; in the second, the left-ear signal was presented diotically to both ears. This diotic presentation of equal signals to both ears can be considered comparable to the perception of a frontal, horizontally non-distributed sound source. The results of the listening experiment revealed no significant differences in the considered SAQI attributes between the two variants, suggesting that representing a human speaker by a point-like sound source with a defined acoustic center is sufficient for plausible binaural rendering.

Keywords: *source width, human voice, virtual acoustic environments*

***Corresponding author**: tobias.weber1@th-koeln.de.*

1. Introduction

A plausible auralization of realistic acoustic scenes in extended reality (XR) requires the accurate representation of natural sound sources, such as human speakers, musical instruments, or everyday objects. In contrast to virtual loudspeakers that are frequently used for spatial audio reproduction, natural sound sources exhibit a wide range of spatial radiation characteristics and are therefore considerably more complex to model. These characteristics include not only the directional (time-variant) radiation patterns, but also their spatial extent. Depending on the distance between the listener and the source, the assumption of a point-like source may not be appropriate, as the physical dimensions or spatial distribution of the source can contribute to its acoustic appearance.

While the aspect of source extent has been investigated in detail for musical instruments [1], [2], studies related to human speakers are rare. In this context, a related important question is whether the spatial extent of a human speaker is perceptually relevant and whether it manifests itself as a perceptible source width. Generally, it has been shown that for human speakers the acoustic sound radiation is phoneme-dependent and shows varying mouth opening sizes [3], [4]. Imbery et al. [5] demonstrated that listeners were able to identify the direction of rotation of a human speaker based on the displacement of the acoustic center. However, as a result of this study, modeling a human speaker by a single sound source located at the acoustic center might be sufficient for a perceptually convincing binaural rendering. For everyday sound objects according to the best knowledge of the authors there are no studies investigating the perceptual relevance of modeling the source extent in virtual acoustic environments.

*Tobias Weber and Hendrik Himmelein contributed equally to this work

In general, natural sound sources differ from ideal point sources due to their their physical dimensions and the distributed excitation of their radiating structures. Even when positioned directly in front of the listener, different radiation positions within such sources result in slightly different propagation paths to the two ears, giving rise to source-related interaural differences. In reverberant environments, these differences are superimposed by additional decorrelation caused by early reflections and reverberation [6]. However, only the source-related interaural differences are directly associated with the spatial extent of the sound source, whereas an ideal frontal point source produces identical signals at both ears.

To investigate the perceptual relevance of source-related interaural differences in the direct sound, the present study follows an approach of systematically manipulating the interaural correlation. Therefore, two conditions of stimuli were created. For the first case, a standard binaural recording was made, containing the naturally occurring interaural differences caused by the spatial extent of the sound source. This stimulus is referred to as dichotic in the following. In contrast, a diotic condition was generated by replacing one ear signal with an identical copy of the other ear signal, resulting in fully correlated ear signals while preserving the overall spectral and temporal characteristics of the original recording. This diotic condition represents an idealized binaural reproduction of a frontally positioned point-like sound source and allows to evaluate the perceptual impact of source-related interaural differences. The evaluation was conducted using the Spatial Audio Quality Inventory (SAQI) [7], which provides a set of attributes for assessing the perceived quality of reproduced virtual acoustic environments.

To further investigate this question across different classes of natural sound sources, we conducted a series of related listening experiments. While a companion study investigates everyday sound sources, such as office and kitchen environments [8], the present paper focuses on human speakers.

2. Method

2.1. Setup

The listening experiments were conducted in the anechoic chamber at TH Köln. Since early reflections and reverberation introduce interaural correlation patterns that are primarily determined by the acoustic environment rather than by the physical dimensions of the sound source, the present study focused exclusively on direct sound.

We defined two fixed human speaker positions, both aligned with the listener position on a single axis. The first position was located 1.5 m in front of the listener, while the second was located 3 m in front of the listener, as shown in Figure 1. By increasing the distance of the source, the interaural differences should decrease, resulting in smaller perceptual differences between the dichotic and diotic signal. An acoustically transparent curtain separated the listener and the human speaker to prevent visual cues from the human speaker. The same setup was used for both the binaural recordings (see Sec. 2.2) and the listening experiment (see Sec. 2.3). During the binaural recordings, a Neumann KU100 artificial head was placed at the listener position used in the listening experiment.

For binaural playback of the recorded samples during the listening experiments, we used AKG K1000 headphones, whose design minimizes possible shadowing effects at the listener's ears [9]. We applied the headphone compensation filters from [10] to equalize the playback. The recordings were reproduced using an RME Babyface Pro FS at a sampling rate of 48 kHz and a buffer size of 256 samples. Since the listening experiment used only static artificial-head recordings and relied on either dichotic or diotic binaural playback, head tracking was not employed, as head rotations cannot be incorporated into this playback concept in a straightforward manner.

The experiment was controlled using custom Max/MSP patches. These patches managed the stimulus order and audio playback, and allowed participants to submit their ratings via a tablet with a graphical user interface (GUI). The audio samples were arranged in the DAW Reaper according to source position, speaker, and dichotic/diotic. Playback in Reaper was controlled from Max/MSP via Open Sound Control (OSC).

2.2. Materials

The test signals consisted of two different sentences constructed using the word matrix of the German matrix sentence test, also known as the Oldenburger Satztest (OLSA)[11]. Prior to the listening experiment, the sentences were spoken by one male and one female speaker and recorded using the KU100 positioned on the defined axis and oriented along the axis, as shown in Figure 1. The height of the KU100 was adjusted so that its interaural center was located 1.29 m above the floor. The human speakers were instructed to maintain their normal speaking level and speaking rate to minimize possible variations during the recordings. Using a handheld sound level meter, the A-weighted sound pressure level $L_{A,eq}$ was measured at the listening position. For a human speaker distance of 3 m, the measured levels during sentence production were 42 dB(A) for the male speaker and 43 dB(A) for the female speaker.

The recordings of both human speakers took place on one single day in ten sessions. The human speakers

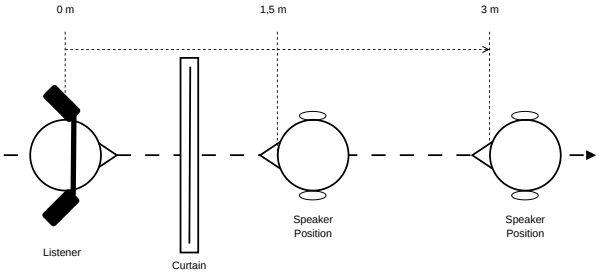

Figure 1: Schematic top view of the experimental setup, showing the defined axis, listener and speaker positions, orientations, and acoustically transparent curtain.

(a) Participant **(b)** Human Speaker

Figure 2: Listening experiment setup with a participant (a) sitting before the curtain and one of the human speakers (b) standing at the 1.5 m position behind the curtain.

alternated after each recording session to account for possible variations in positioning and speaking level. In each session, the sentences were recorded from both predefined positions. In total, there were two human speakers (one female and one male), who spoke two different sentences from two different positions. Both human speakers were present during all listening experiments and spoke the same sentences as references from the corresponding positions.

For each binaural recording, a diotic signal was created by replacing the right channel with a copy of the left channel, resulting in two variants of each recording: the original dichotic signal, in which the left and right channels differed, and the diotic signal, in which both channels contained identical signals.

2.3. Procedure

Before the start of each listening experiment, participants received detailed instructions on the individual SAQI attributes so that naive listeners could understand and evaluate them appropriately. They were then led into the anechoic chamber and seated on a swivel chair in front of the acoustically transparent curtain, as shown in Figure 2. The height of the chair was adjusted so that each participant's interaural center matched the height of the KU100 used for the recordings. Participants were then provided with headphones and a tablet for rating the SAQI attributes using Max/MSP. Once all remaining questions had been answered, both human speaker went behind the curtain and the experiment started. During the experiments, participants were instructed to keep their heads facing forward while the stimuli were presented and to avoid additional head movements, except when using the tablet to submit their ratings.

The experiment consisted of 16 trials in total. In each trial, participants rated the headphone playback of a stimulus in comparison with the reference, i.e., the corresponding human speaker. The order of the individual stimulus conditions was randomized in Max/MSP at the beginning of the experiment. Participants could listen to the sentence spoken by the human speaker "A" and to the corresponding headphone playback sample "B" as often as desired by pressing the "A" or "B" button. The human speaker and the headphone playback always corresponded to the same sentence and position, either 1.5 m or 3 m from the listener. As described in Sec. 2.2, there were ten headphone samples for each condition, which were selected at random when Button 'B" was pressed for headphone playback.

For each trial, participants were asked to rate eight SAQI attributes using sliders: "difference", "distance", "source width", "externalization", "localization", "spatial disintegration", "naturalness", and "degree of liking". These SAQI attributes were selected to enable the assessment not only for human speakers, but also for other natural sound sources [8] allowing for direct comparison across studies. In each trial, participants first rated the attribute "difference" on a uniform scale to assess whether they perceived any difference between the reference and the headphone playback. Only after submitting this rating were the following SAQI attributes displayed: "distance", "source width", and "externalization". The remaining four attributes, "localization", "spatial disintegration", "naturalness", and "degree of liking", were rated on the subsequent page. The rating procedure followed the SAQI test design specified by Lindau et al. [7].

2.4. Participants

A total of 19 participants (two female, one diverse, and 16 male) took part in the listening experiments, aged 19–57 years (mean age: 31.05 years). Two participants self-reported a slight low-frequency hearing loss, while the remaining participants reported normal hearing. Nine participants were naive listeners or had participated in at most one previous listening experiment, whereas the other ten participants were considered expert listeners.

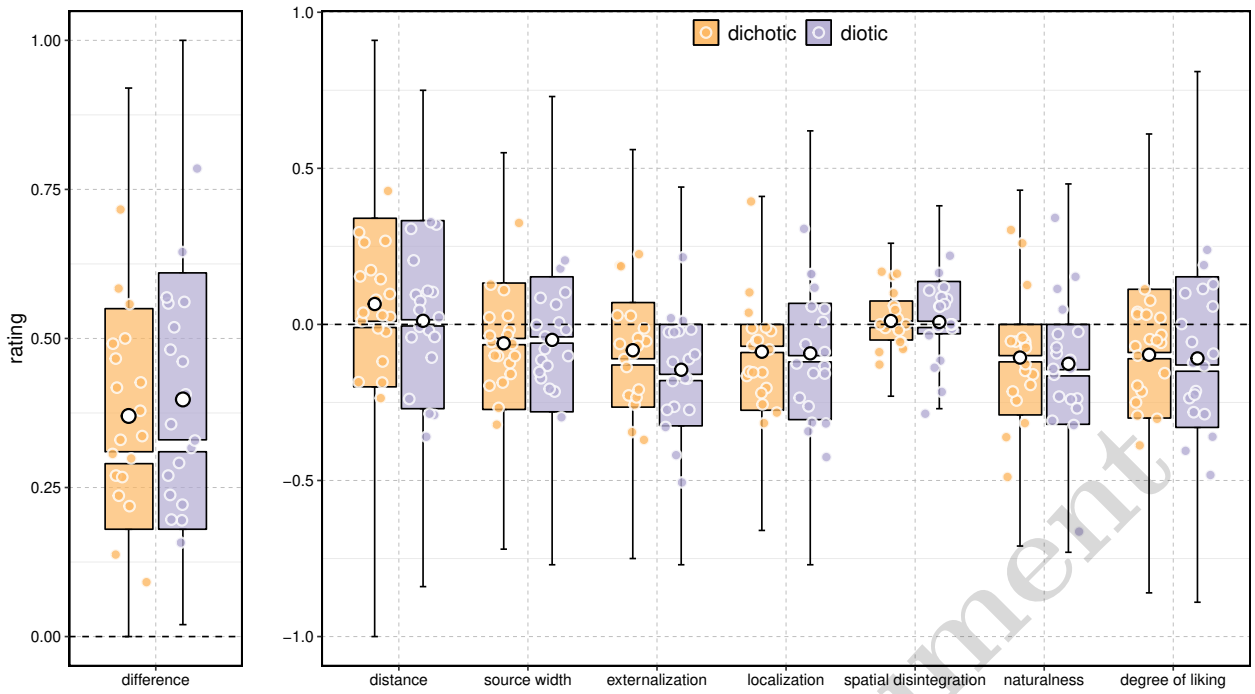

Figure 3: Overall ratings per SAQI attribute including the mean (white dot), the median (white bar) with interquatil ranges (colored boxes) averaged per participant.

2.5. Data Analysis

Data was analyzed using multivariate analysis of variance (MANOVA). A MANOVA was conducted for all evaluated parameters simultaneously to test the effects of *dichotic/diotic* and *position*, while pooling across the *speech material* (speaker and sentence). All fixed effects were treated as categorical factors, and the model included their full factorial interaction. Multivariate significance was assessed using Pillai's trace.

Following this, univariate ANOVAs were subsequently conducted for each dependent variable separately using the same fixed-effects structure. Effect size was quantified using R^2 and $R^2_{adjusted}$ from a linear model fit with identical structure. This was done using the performance package [12]. Post-hoc pairwise comparisons with Tukey correction were conducted for each evaluated parameter using estimated marginal means, separately for every main and interaction effect, utilizing the emmeans package [13].

Subsequently, a multivariate Bayesian regression model was fitted using the brms package [14]. All evaluated parameters were modeled together for the described fixed effects with a random intercept for participant. To this end, all parameters were rescaled to the open interval (0,1) and modeled using a Beta likelihood. Default priors were used. Bayes factors were calculated by testing the model against a null model excluding the target factor dichotic/diotic, estimated via bridge sampling [15].

3. Results

Figure 3 shows the overall differences between the dichotic and diotic representation averaged over all positions and speech material. General perceived differences is depicted to the left ranging from 0 (no difference) to 1 (completely different) while other SAQI attributes are shown to the right ranging from -1 to +1. However, results show only a minor shift towards an overall higher rated differences for the diotic case and almost no differences across the considered SAQI attributes.

To further analyze possible differences a three-way multivariate ANOVA with dichotic/diotic, position and speech material as independent variables was conducted for the overall differences and separately for all investigated attributes. The initial MANOVA showed no significant effects for the main effects dichotic/diotic ($p = 0.77$), position ($p = 0.19$) and speech material ($p = 0.89$), nor for the interaction effects dichotic/diotic x position ($p = 0.95$), dichotic/diotic x speech material ($p = 0.66$), position x speech material ($p = 0.40$) and dichotic/diotic x position x speech material ($p = 0.75$).

Similarly, almost no significant results were found evaluating each SAQI attribute independently ($R^2 \leq 0.06$ for all attributes). Only "distance" showed a significant difference across positions, $F(1, 288) = 7.11, p = 0.008$ with $R^2 = 0.048$.

Following a subsequent bayesian model was fitted as described in Sec. 2.5. This comparison yielded a Bayes factor of $BF_{10} > 100$ in favor of the null model, indicating decisive evidence [15] against an effect of the dichotic/diotic manipulation, consistent with the frequentist results reported above.

4. Discussion

The results indicate that for human speakers, replacing the dichotic binaural reproduction, which preserved the naturally occurring differences between the left and right ear signals, with a diotic reproduction using identical ear signals did not result in perceptible changes in source width or other relevant SAQI attributes. The results are consistent with the assumption of a point-like sound source with a defined acoustic center, or more generally, a sound source without a relevant horizontal extent which seems to be well fulfilled for human speakers.

Several aspects of human voice radiation that could potentially influence the perceived spatial characteristics of a speaker, including varying mouth opening sizes [3], [4], asymmetries in the radiation pattern [16] and the vertical extent of the sound source caused by effects such as torso reflections [17], [18] did not have a relevant impact under the investigated conditions, as they do not appear to introduce a relevant interaural decorrelation of the ear signals.

Future studies will investigate whether the observed findings generalize to other classes of natural sound sources. A companion study that currently is in preparation [8] indicates that source-related interaural differences are perceptually relevant for many everyday sound sources which suggests that the relevance of the investigated attributes depends on the type and the physical structure of the sound source. Furthermore, it needs to be investigated how the results are affected by non-frontal source positions and head rotations. Finally, the perceptual impact of room reflections and reverberation, which introduce additional interaural decorrelation, should be investigated. Overall, this study indicates that, for human speakers under anechoic conditions, a point-like source representation with a defined acoustic center is sufficient for plausible binaural rendering.

5. Acknowledgments

The authors gratefully acknowledge Lyanna Koch for acting as speaker in this study. Moreover, great thanks to all participants in the listening experiments.
The work carried out in this study has been funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the projects "Investigations on the perception of dynamic directivity of human voice" (project ID 529603294) and "Investigations into the plausible presentation of natural sound sources in virtual acoustic environments" (project ID 572159676). We appreciate the support.
This study was approved by the Commission on Responsibility in Science Ethics Committee at TH Cologne (Application number: THK-2023-0006).

References

[1] D. Ackermann, F. Brinkmann, and S. Weinzierl, "A Database with Directivities of Musical Instruments," *Journal of the Audio Engineering Society*, vol. 72, no. 3, pp. 170–179, Mar. 2024, ISSN: 15494950. DOI: 10.17743/jaes.2022.0128.

[2] D. Ackermann, F. Brinkmann, and S. Weinzierl, "Musical instruments as dynamic sound sources," *The Journal of the Acoustical Society of America*, vol. 155, no. 4, pp. 2302–2313, Apr. 2024, ISSN: 0001-4966. DOI: 10.1121/10.0025463.

[3] C. Pörschmann and J. M. Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4553–4564, Jun. 2021, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/10.0005401.

[4] C. Pörschmann and J. M. Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns II: Various groups of phonemes," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 179–190, Jan. 2023, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/10.0016821.

[5] C. Imbery, S. Franz, S. Van De Par, and J. Bitzer, "Auditory Facing Angle Perception: The Effect of Different Source Positions in a Real and an Anechoic Environment," *Acta Acustica united with Acustica*, vol. 105, no. 3, pp. 492–505, May 2019, ISSN: 1610-1928. DOI: 10.3813/AAA.919331.

[6] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 255–265, Jul. 1998, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.423955.

[7] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994, Sep. 2014, ISSN: 16101928. DOI: 10.3813/AAA.918778.

[8] H. Himmelein, T. Weber, and C. Pörschmann, "Influence of interaural correlation on perceived source width in binaural rendering of natural sound sources," (in preparation).

[9] C. Pörschmann, J. M. Arend, and R. Gillioz, "How wearing headgear affects measured head-related transfer functions," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, 2019, pp. 49–54.

[10] B. Bernschütz, "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, Merano, Italy, 2013, pp. 592–595.

[11] K. Wagener, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests," *Zeitschrift für Audiologie*, no. 38, p. 32, 1999.

[12] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski, "Performance: An R package for assessment, comparison and testing of statistical models," *Journal of Open Source Software*, vol. 6, no. 60, p. 3139, 2021-03-23. DOI: 10.21105/joss.03139.

[13] R. V. Lenth and J. Piaskowski, *Emmeans: Estimated marginal means, aka least-squares means*, R package version 2.0.3, 2026. [Online]. Available: https://CRAN.R-project.org/package=emmeans.

[14] P.-C. Burkner, "Brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017. DOI: 10.18637/jss.v080.i01.

[15] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995. DOI: 10.1080/01621459.1995.10476572.

[16] C. Pörschmann and J. M. Arend, "Phoneme dependence of horizontal asymmetries in voice directivity," *JASA Express Letters*, vol. 4, no. 2, p. 025 205, 2024, ISSN: 2691-1191. DOI: 10.1121/10.0024878.

[17] R. Blandin, J. Geng, and P. Birkholz, "Investigation of the influence of the torso, lips and vocal tract configuration on speech directivity using measurements from a custom head and torso simulator," *Acta Acustica*, vol. 7, p. 39, 2023, ISSN: 2681-4617. DOI: 10.1051/aacus/2023035.

[18] C. Pörschmann and J. M. Arend, "On the impact of downward-directed human voice radiation on ground reflections," *Acta Acustica*, vol. 8, p. 12, 2024, ISSN: 2681-4617. DOI: 10.1051/aacus/2024002.