Direct-to-Reverberant Energy Ratio Estimation and Extrapolation from Own Speech

Nils Meyer-Kahlen*
Dpt. of Information and Communications Engineering
Aalto University
Espoo, Finland
nils.meyer-kahlen@aalto.fi

Thomas Deppisch*
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden
thomas.deppisch@chalmers.se

***Abstract*—Accurately characterizing a user's acoustic environment is essential for creating virtual sound sources in augmented reality that blend seamlessly into the real environment. The acoustic parameters of an environment can be calculated from a room impulse response (RIR) and the authors recently presented a method to blindly estimate RIRs from speech signals captured with a head-worn microphone array. The approach uses either speech from a distant speaker or own speech from the person wearing the array on their head. While both variants provide reliable reverberation time estimates, direct-to-reverberant energy ratio (DRR) estimates from the user's own speech deviate significantly from the expected DRR of a distant virtual source due to the higher direct sound level. This study investigates the feasibility of extrapolating DRR values from own speech to predict DRRs of distant sources. The approach relies on two acoustic assumptions: (i), the mouth-to-array transfer paths do not change significantly between users and, (ii), a homogeneous reverberant field. Our findings show that the assumptions hold above the Schröder frequency and in sufficiently reverberant conditions. Average DRR extrapolation errors are below 2 dB at mid frequencies when using mouth simulator measurements and around 3 dB with actual speech recordings.**

***Index Terms*—Augmented Reality, Direct-to-Reverberant Energy Ratio, Room Acoustics, Room Impulse Response**

I. Introduction

To seamlessly integrate virtual sound sources into real acoustic scenes in mixed or augmented reality (MR/AR) applications, it is essential to characterize the acoustic properties of the user's environment [1], [2]. Specifically, AR rendering techniques aim to match the reverberation time (RT) and direct-to-reverberant energy ratio (DRR) of rendered virtual sources to the acoustic conditions of the real environment. A variety of methods exist to blindly estimate these parameters [3]–[9] or to estimate RIRs from which the parameters can be computed [10]–[14].

Head-worn devices, such as smart glasses or MR headsets, have become a key medium for delivering AR or MR experiences. Since such devices typically comprise microphone arrays, this has introduced a new context for estimating RT and DRR: leveraging the user's speech wearing the microphone array.

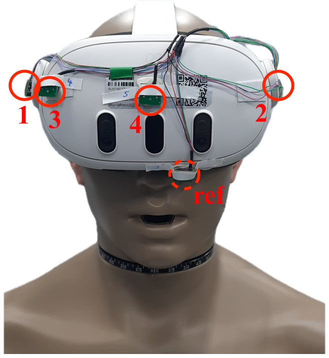

Fig. 1: The KEMAR head and torso simulator incl. mouth simulator wearing a Meta Quest 3 with a microphone array comprising four target microphones (solid red circles) and one reference microphone (dashed red circle).

In our previous work [14], we compared RT and DRR derived from RIR estimates using speech from a distant speaker (*far voice*) versus the user's own speech (*own voice*). With the aim of rendering distant virtual sound sources in mind, estimation results were compared against a reference measured from a distant source for both cases. The RT estimation errors were comparable for both approaches, suggesting that RT estimates from the own voice can be used to render distant virtual sources. However, DRR estimation errors from the own voice approach were significantly higher than from distant sources. This discrepancy is expected, as the DRR for the user's own voice, close to the array, is naturally much higher than that of a far-field source.

In this contribution, we analyze whether DRR estimates from own voice can be extrapolated to match DRRs of distant sources using the head-worn microphone array in Fig. 1. The result is important for any practical AR estimation and rendering approach that is based on room acoustic estimation from the user's own voice. We show that the feasibility depends on two basic acoustic assumptions: (i), the mouth-to-array transfer path must not change significantly when the array is worn by different users and in different spaces, and, (ii), the reverberant field must be approximately homogeneous, i.e., have equal reverberant energy at any position in the room.

II. Background: Blind Estimation of the DRR

Any practical speech-based DRR extrapolation method must be based on DRR estimates and we employ the approach

*Both authors contributed equally to this work. Thomas Deppisch thanks Reality Labs Research for funding his research.

from [14] to blindly estimate RIRs, from which DRRs can be extracted.

In the frequency domain, microphone signals $X(\omega)$ due to the speech signal $S(\omega)$ are described by a multiplication with the room transfer function $H_{\mathrm{near}}(\omega)$ between the speaker's mouth and a microphone,

$$X(\omega) = H_{\mathrm{near}}(\omega)\, S(\omega)\,. \tag{1}$$

If an estimate of the dry speech signal $\tilde{S}(\omega)$ is available, a transfer function estimate can be obtained from a Wiener filter,

$$\tilde{H}_{\mathrm{near}}(\omega) = \frac{\mathrm{E}\{\tilde{S}^{*}(\omega)X(\omega)\}}{\mathrm{E}\{\tilde{S}^{*}(\omega)\tilde{S}(\omega)\} + \delta}\,, \tag{2}$$

where δ is a small regularization constant. The signal $\tilde{S}(\omega)$, referred to as the pseudo reference signal in [14], is obtained from multichannel array signals using beamforming and dereverberation. However, as shown in that study, dereverberation is unnecessary when estimating RIRs from the user's own speech, given the close proximity of the microphones to the speech source. In the scenarios presented here, we found that using the signal from the microphone closest to the mouth, see Fig. 1, as a pseudo reference, without applying dereverberation or beamforming, yielded results comparable to those obtained with beamforming. Therefore, all results presented in this study are based on this simplified approach.

The objective now is to estimate the DRR from $\tilde{H}_{\mathrm{near}}$ and adjust it to match the DRR of the room transfer function measured at a given distance H_{far}. The following section introduces the necessary definitions and assumptions.

III. Method: DRR Extrapolation

Room impulse responses $h(t)$ can be modeled as being composed of the direct sound $d(t)$ and the reverberant sound $r(t)$ containing early reflections as well as late reverberation,

$$h(t) = d(t) + r(t)\,. \tag{3}$$

The DRR is defined as the ratio of their energies. Assuming that the direct sound and the reverberant sound starting from the first reflection are disjunct in time, the DRR can in practice be determined as

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{t=0}^{\infty} d^2(t)}{\sum_{t=0}^{\infty} r^2(t)}\right) \approx 10\log_{10}\left(\frac{\sum_{t=0}^{t_{\mathrm{d}}} h^2(t)}{\sum_{t=t_{\mathrm{d}}}^{\infty} h^2(t)}\right) \tag{4}$$

where t_{d} represents the time when the direct sound ends and the first reflection begins.

The corresponding transfer paths are in the following described in the frequency domain and denoted by capital letters $D(\omega)$ and $R(\omega)$, representing their time-domain counterparts $d(t)$ and $r(t)$. The frequency-dependent DRR is then defined as

$$\mathrm{DRR}(\omega) = 10\log_{10}\left(\frac{|D(\omega)|^2}{|R(\omega)|^2}\right). \tag{5}$$

For direct sound extrapolation, the relative transfer path $D_{\mathrm{n2f}}(\omega)$ of the direct sound from the own mouth to a far speaker position at a reference distance r_{ref} needs to be determined. Here, it is measured by an anechoic transfer function measurement between a mouth simulator of a dummy head wearing the microphone array and a distant dummy head's mouth simulator to the microphone array, see Fig. 2. The two direct sound paths are denoted as $D_{\mathrm{near}}(\omega)$ and $D_{\mathrm{far}}(\omega)$, respectively, and the transfer path from the near to far position is determined as

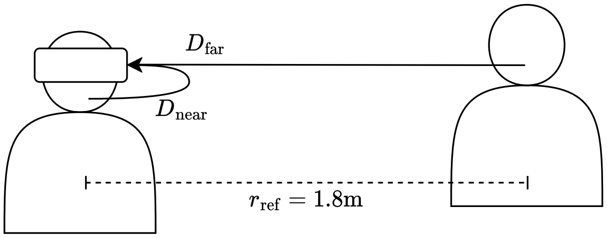

Fig. 2: Near and far direct sound path for one of the array microphones.

$$D_{\mathrm{n2f}}(\omega) = \frac{D_{\mathrm{far}}(\omega)}{D_{\mathrm{near}}(\omega)}\,. \tag{6}$$

One such transfer function is determined for each array microphone beside the reference microphone. With assumption (i), stating that $D_{\mathrm{n2f}}(\omega)$ is robust to the placement of the array on different user's heads and independent of the room in which the user is located, the direct sound transfer function estimate at the near position in a room $\tilde{D}_{\mathrm{near}}(\omega)$ can be extrapolated to any other position at distance r, azimuth ϕ and elevation θ, by taking into account the distance attenuation relative to the reference position from the anechoic measurement r_{ref}, and the directivity pattern $\Gamma(\omega,\phi,\theta)$ of the mouth, relative to the direction used to determine $D_{\mathrm{n2f}}(\omega)$,

$$\tilde{D}_{\mathrm{far}}(\omega,r,\phi,\theta) = \frac{r_{\mathrm{ref}}}{r}\,\Gamma(\omega,\phi,\theta)\, D_{\mathrm{n2f}}(\omega)\, \tilde{D}_{\mathrm{near}}(\omega)\,. \tag{7}$$

Assuming, (ii), a homogeneous reverberant field, the reverberant energy is the same everywhere in a room. If this assumption holds, an estimate of the reverberant energy at the near position $\tilde{R}_{\mathrm{near}}(\omega)$ can directly be utilized as estimate for the reverberant energy at any distant position,

$$|\tilde{R}_{\mathrm{far}}(\omega)|^2 \approx |\tilde{R}_{\mathrm{near}}(\omega)|^2\,. \tag{8}$$

In a given acoustic environment, the frequency dependent DRR of the extrapolated estimate can now be calculated from the extrapolated energies $|\tilde{D}_{\mathrm{far}}(\omega)|^2$ and $|\tilde{R}_{\mathrm{far}}(\omega)|^2$ using (5). Sec. V-A investigates if the assumptions hold in practice.

IV. Experiment Setup

A. Anechoic Measurements

The anechoic measurement to determine $D_{\mathrm{near}}(\omega)$ was conducted using a KEMAR head and torso simulator wearing the head-mounted array, with its mouth simulator serving as the sound source (see Fig. 1). To determine $D_{\mathrm{far}}(\omega)$, a separate measurement was performed with the KEMAR and its mouth simulator positioned at a reference distance of $r_{\mathrm{ref}} = 1.8$ m, while the head-mounted array was worn by a Cortex Instruments MK1 head and torso simulator.

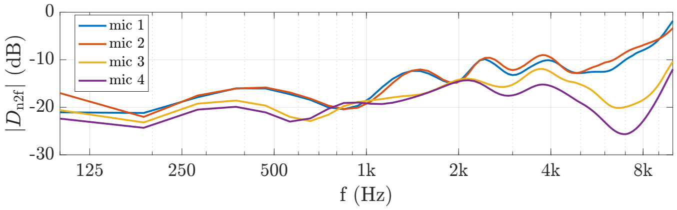

Fig. 3: Magnitude of $D_{\mathrm{n2f}}(\omega)$ for the extrapolation of the direct sound energy from the near to the distant reference position.

Fig. 3 displays the third-octave-smoothed magnitude of $D_{\mathrm{n2f}}(\omega)$ obtained from these measurements. Due to the directivity of the mouth simulator and the positioning of the microphones on the array, microphones 1 and 2 require a stronger boost for extrapolation to the reference distance compared to microphones 3 and 4, especially at high frequencies.

B. *Variable-Acoustic Measurements and Recordings*

The next set of measurements was conducted in the variable acoustics room "Arni" at the Aalto Acoustics lab. It has 55 acoustic panels distributed along the four walls that can be individually set to be either absorbing or reflecting. Six room settings were used, summarized in Tab. I.

First, KEMAR measurements were taken, similar to the anchoic measurements. For the near measurements, the array was positioned on the KEMAR while it produced a sweep through its mouth simulator. Then, the KEMAR was placed at four distances ($1\,\mathrm{m}$, $2\,\mathrm{m}$, $3\,\mathrm{m}$, and $4\,\mathrm{m}$) directly in front of the Cortex head wearing the microphone array. Both heads were oriented toward each other during the measurements.

Fig. 4 shows the ground truth DRRs for different source distances, as well as the DRRs from the reference microphone at the *near* position (black markers), where the KEMAR head was wearing the microphone array. The DRR values from the reference microphone measurements are clearly the highest, ranging from 25 to $30\,\mathrm{dB}$. Overall, the DRRs exhibit a clear decreasing trend as the source-receiver distance increases, approximately following the $1/r$ inverse-distance relation. Excluding the reference microphone, the dataset covers DRRs ranging from approximately -10 to $20\,\mathrm{dB}$.

In addition to the measurements with KEMAR, three female and three male participants wearing the microphone array were recorded in all six room configurations. Each participant repeated three sentences from the first set of Harvard sentences [15] three times.

C. *Processing*

Blind RIR estimation from own speech was performed using $6\,\mathrm{s}$ of convolution-based speech signals (generated from measured RIRs with a KEMAR mouth simulator) and $6\,\mathrm{s}$ of recorded speech from six participants wearing the microphone array. The closest microphone to the mouth (see Fig. 1) served as the reference, providing $\tilde{S}(\omega)$, while the other four were target microphones for RIR estimation. Signals were recorded and processed at $48\,\mathrm{kHz}$, with the Wiener filter's expected value in (2) estimated via $1\,\mathrm{s}$ STFT blocks (hop size: $43\,\mathrm{ms}$). The regularization constant was $\delta = 10^{-4}$, and RIR estimates were truncated to half the block length to remove noise floor and acausal components.

TABLE I: Descriptions of six different room configurations based on acoustic panel settings, along with their respective broadband reverberation times.

Room Setting	**Description**	**Reverb Time (s)**
100% Dry	All panels are set to their absorbing (dry) side.	0.27
100% Reverberant	All panels are set to their reflecting (wet) side.	0.81
25% Dry	25% of the panels are absorbing, the remaining 75% reflecting, with both types evenly distributed.	0.49
50% Dry	Half of the panels are absorbing, the other half reflecting, with both types evenly distributed.	0.35
Dead-End-Live-End (DELE)	One half of the room is entirely absorbing, while the other half is fully reflective.	0.41
Live-End-Dead-End (LEDE)	The reverse of DELE: one half is reflective, while the other is fully absorbing.	0.36

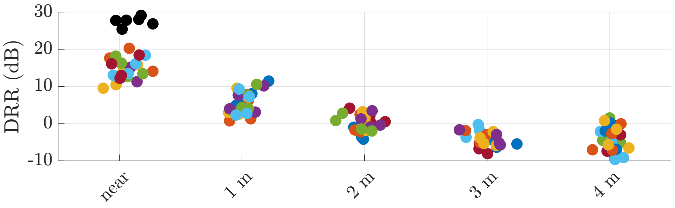

Fig. 4: DRRs from the dataset at different distances from the microphone array. Black markers at the *near* position show the DRR at the reference microphone. Other markers with the same color illustrate DRRs from different microphones in the same room acoustic condition.

RIRs (both ground truth and estimated) were split into direct and reverberant parts. For near measurements (KEMAR or human-worn array), the split occurred $7.7\,\mathrm{ms}$ after the direct sound peak, before the first floor reflection. For distant speech, it was $2.1\,\mathrm{ms}$ to exclude the earliest floor reflection (at the $4\,\mathrm{m}$ distant position). The *near* direct sound was extrapolated to all four far positions using (7), with $D_{\mathrm{n2f}}(\omega)$ smoothed in third-octave bands and applied as a minimum-phase filter.

The reverberant parts had noise floors removed using the Lundeby method [16] from [17]. DRRs were computed in the frequency domain according to (5) and smoothed in third-octave bands.

V. Results

A. *Validity of Assumptions*

In the following, we analyze the validity of the two assumptions described above by separately analyzing the extrapolation results for direct and reverberant part.

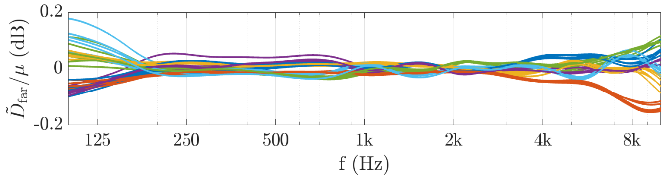

Fig. 5: Variation in extrapolated direct sound per human subject. (Same color indicates same subject in different room conditions, averaged over positions and microphones).

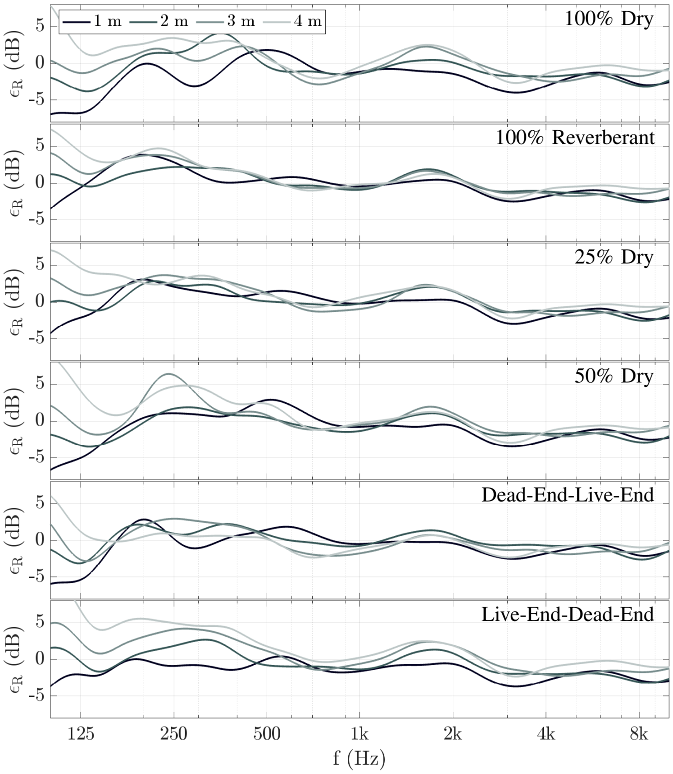

Fig. 6: Signed error of the reverberant energy ϵ_{R} at different distances, averaged over microphones.

Assumption (i), the invariance of the mouth-to-array transfer function, is tested using the RIR estimates from the human recordings. A somewhat different fit across participants can be expected to cause the direct path to vary. Results for the extrapolated direct sound normalized by the mean over all direct sound estimates μ are visualized in Fig. 5, which shows the variation of the extrapolated direct sound is negligible and that assumption (i) is clearly fulfilled.

Next, we check assumption (ii), which is the homogeneity of the reverberant field. Here, we use KEMAR measurements, comparing the reverberant energy at the distant and the near positions. Fig. 6 shows the errors of the extrapolated reverberant energy separately for each room condition. The errors for the 100% reverberant, 25% dry, and 50% dry conditions are similar. Between 1 kHz and 2 kHz, there is a tendency for larger positive errors at greater distances. The largest errors occur in the driest condition (100% dry) and the Live-End-Dead-End (LEDE) condition, where reverberant energy is increasingly overestimated with distance. This can be explained by the fact that, in the less reverberant field of the dry condition, homogeneity of the sound field from the reverberant sound is not reached. Similarly, in the LEDE condition, less energy remains in the room, as a significant portion of the source energy is immediately absorbed by the wall opposite the source.

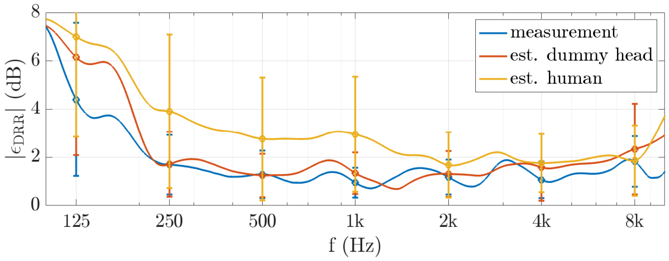

Fig. 7: Mean and standard deviation of absolute DRR error ϵ_{DRR} for measurement-based extrapolation, extrapolation using speech from a dummy head with mouth simulator, and extrapolation using speech from human subjects.

As expected, larger errors are also observed at low frequencies around 100 Hz. Here, the response is likely dominated by strong room modes, as also indicated by the Schröder frequency which lies between 90 and 160 Hz for the different conditions, resulting in noticeable position-dependent pressure variations.

B. *DRR Extrapolation Performance*

Fig. 7 compares the mean absolute DRR error, ϵ_{DRR}, obtained by averaging the results from the four target microphones in the array, the four extrapolation distances, and the six acoustic conditions. For the blind estimation results, the errors are additionally averaged over six different speech samples.

The absolute DRR errors from the measurement-based extrapolation remain below 2 dB on average between approximately 200 Hz and 10 kHz but increase to around 7 dB at 100 Hz. This indicates that errors resulting from violations of the homogeneity assumption are in practice manageable above 200 Hz.

A similar trend is observed for blind estimation using speech convolved with mouth simulator impulse responses. The errors only slightly increase, demonstrating that the limited spectral content of speech does not significantly impact the accuracy of the estimation.

However, the most notable errors occur when DRR estimation is performed using real human speakers wearing the array. In this case, errors increase to approximately 3 dB at mid frequencies and around 4 dB at low frequencies near 200 Hz.

C. *Discussion*

While assumption (i) seems to hold well, assumption (ii), which states that the reverberant field is homogeneous, is valid only under sufficiently reverberant conditions and above the Schröder frequency. One approach to improving the estimated reverberant energy is incorporating alternative acoustic models. A natural choice would be to apply Barron's *revised* *theory* [18] of energy relationships. This theory predicts a decay in reverberant energy as a function of distance, with greater decay occurring in smaller and less reverberant rooms. While it would successfully explain the data between 1 kHz and 2 kHz in the dry condition, its application to the present dataset resulted in values that were too low at other frequencies. Moreover, it does not account for the difference between the DELE and LEDE conditions, which have similar reverberation times and room volumes. Also, using it would require knowledge of the room volume.

In MR/AR applications, virtual sources would be rendered based on the DRR estimates. Thus, ultimately, a perceptual evaluation is necessary. To gain an initial impression of the perceptual significance of the errors, the results can be compared to the just noticeable difference (JND). According to [19], the JND for DRR follows a U-shaped pattern, with the threshold depending on the absolute DRR. It is approximately 2 dB for absolute DRRs close to 0 dB and increases for both lower and higher absolute DRRs.

This suggests that, in direct comparison, the DRR errors observed in practical speech-based estimates are likely to be perceptible, at least in the 2 m condition, where the absolut DRR is around 0 dB (see Fig. 4). However, in a real-world AR application, a direct comparison between one's own voice and its virtual counterpart is not possible. When comparing one's own voice to the rendered version, larger errors may be tolerable. Similarly, when comparing renderings of distant sources emitting different signals, the detection threshold for DRR differences is likely to increase.

Following a more radical thought, perfect DRR extrapolation might not even be desired. It might be preferable to render distant sources with characteristics similar to the room excitation caused by the user's own voice, particularly at low frequencies. While listeners likely do not expect a distant voice to have the same DRR as their own, it is less clear whether they can form accurate expectations about the modal excitation a distant source would produce compared to their own speech, i.e., the larger errors at low frequencies might turn out to even be desirable. In-situ listening tests are required to address these questions.

VI. Conclusion

This study investigated the feasibility of extrapolating DRR values from RIR estimates based on own speech from a person wearing a head-mounted microphone array, so that it matches the DRR of sound sources at different distances from the user. Our findings indicate that the extrapolation, based on the assumption of a homogeneous reverberant field, is effective in environments that are sufficiently reverberant and at frequencies above the Schröder frequency and that variations in transfer paths between user's mouths and the array due to different fittings are not an issue in practice. Absolute DRR extrapolation errors were below 2 dB on average at mid frequencies when extrapolating from mouth simulator measurements and approximately 3 dB when using actual speech recordings from subjects wearing the array.

Acknowledgements

We thank the volunteers who recorded sentences, and Raimundo Gonzalez for providing the microphone array.

References

[1] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends in Hearing*, vol. 26, p. 1–22, 2022.

[2] N. Meyer-Kahlen, S. V. Amengual Garí, S. J. Schlecht, and T. Lokki, "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility and Authenticity," *J. Audio Eng. Soc.*, vol. 72, no. 11, pp. 797–812, 2024.

[3] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of Room Acoustic Parameters: The ACE Challenge," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[4] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," *IEEE Int. Workshop on Acoustic Signal Enhancement*, pp. 136–140, 2018.

[5] A. Perez-Lopez, A. Politis, and E. Gomez, "Blind reverberation time estimation from ambisonic recordings," in *IEEE 22nd International Workshop on Multimedia Signal Processing*, 2020, pp. 1–6.

[6] W. Mack, S. Deng, and E. A. Habets, "Single-channel blind direct-to-reverberation ratio estimation using masking," in *Proceedings of INTERSPEECH*, 2020, pp. 5066–5070.

[7] P. Calamia, N. Balsam, and P. Robinson, "Blind estimation of the direct-to-reverberant ratio using a beta distribution fit to binaural coherence," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. EL359–EL364, 2020.

[8] S. Saini and J. Peissig, "Blind Room Acoustic Parameters Estimation Using Mobile Audio Transformer," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023.

[9] C. Wang, M. Jia, M. Li, C. Bao, and W. Jin, "Exploring the power of pure attention mechanisms in blind room parameter estimation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 23, p. 1–18, 2024.

[10] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021, pp. 221–225.

[11] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha, and P. Calamia, "Towards Improved Room Impulse Response Estimation for Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[12] S. Lee, H. S. Choi, and K. Lee, "Yet Another Generative Model for Room Impulse Response Estimation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023, pp. 1–5.

[13] J.-M. Lemercier, E. Moliner, S. Welker, V. Välimäki, and T. Gerkmann, "Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models," pp. 1–13, 2024. [Online]. Available: http://arxiv.org/abs/2408.07472

[14] T. Deppisch, N. Meyer-Kahlen, and S. V. Amengual Garí, "Blind Identification of Binaural Room Impulse Responses from Smart Glasses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, p. 4052–4065, 2024.

[15] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969, appendix C.

[16] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorländer, "Uncertainties of measurements in room acoustics," *Acta Acustica united with Acustica*, vol. 81, no. 4, pp. 344–355, 1995.

[17] M. Berzborn, R. Bomhardt, and J. Klein, "The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing," 2017.

[18] M. Barron and L. Lee, "Energy relations in concert auditoriums. I," *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 618–628, 1988.

[19] 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.