FORUM ACUSTICUM EURONOISE 2025

BINAURAL LOCALIZATION MODEL FOR SPEECH IN NOISE

Vikas Tokala[1*] **Eric Grinstein**[1] **Rory Brooks**[1]
Mike Brookes[1] **Simon Doclo**[2] **Jesper Jensen**[3,4] **Patrick A. Naylor**[1]
[1] Dept. of Electrical and Electronic Engineering, Imperial College London, UK
[2] Dept. of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Germany
[3] Demant A/S, Smørum, Denmark
[4] Dept. of Electronic Systems, Aalborg University, Denmark

ABSTRACT

Binaural acoustic source localization is important to human listeners for spatial awareness, communication and safety. In this paper, an end-to-end binaural localization model for speech in noise is presented. A lightweight convolutional recurrent network that localizes sound in the frontal azimuthal plane for noisy reverberant binaural signals is introduced. The model incorporates additive internal ear noise to represent the frequency-dependent hearing threshold of a typical listener. The localization performance of the model is compared with the steered response power algorithm, and the use of the model as a measure of interaural cue preservation for binaural speech enhancement methods is studied. A listening test was performed to compare the performance of the model with human localization of speech in noisy conditions.

Keywords: *Binaural source localization, reverberation, human hearing, interaural cues, spatial hearing*

1. INTRODUCTION

Binaural localization has garnered significant attention in the field of Computational Auditory Scene Analysis (CASA), which is influenced by principles underlying the perceptual organization of sound by human listeners. The two primary cues for sound localization are the Interaural Time Differences (ITD), also known as the time difference of arrival, and the Interaural Level Difference (ILD), which arises due to the influence of the head, torso, and outer ear. Differences between localization methods often stem from varying assumptions about environmental factors such as sound propagation, background noise, and microphone configuration. Localizing sound sources using binaural input in noise and reverberation is a challenging problem with important applications in hearing aids, spatial sound reproduction, and mobile robotics.

It is well established that the noise and reverberation in typical listening environments can mask signals and negatively affect both binaural and monaural spectral cues, leading to reduced sound localization accuracy and speech comprehension even for individuals with normal hearing [1–3]. Research has shown that localization accuracy declines as the Signal-to-Noise Ratio (SNR) decreases. For instance, [1] studied three normal hearing listeners who were asked to localize broadband click trains in an anechoic chamber under one quiet and nine noisy conditions with SNRs ranging from -13 to +14 dB. Their findings revealed that localization accuracy was poorest in the lateral horizontal plane and began to deteriorate at SNRs below +8 dB. Similarly, [2] investigated the effect of SNR on localization ability in normal hearing listeners, finding that typical environments characterized by both noise and reverberation can further degrade localization cues and impair performance. In [4], it is suggested that the combined effects of noise and reverberation could further reduce localization accuracy. A well-known method for localisation using ITD estimation is the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) approach, which assumes ideal single-path propagation.

*__Corresponding author__: v.tokala@imperial.ac.uk.

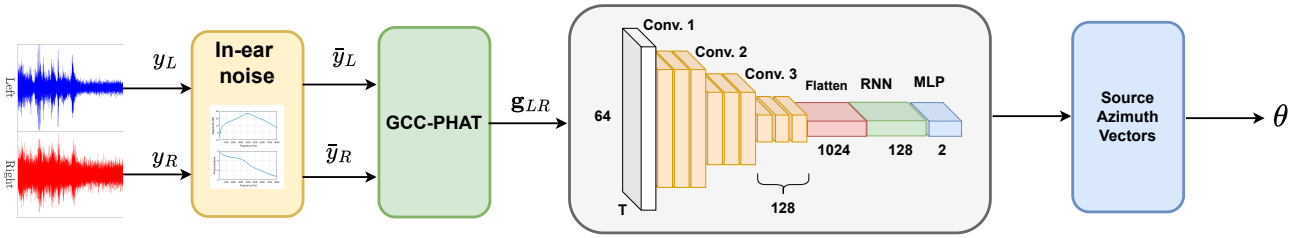

Figure 1: Block diagram of the model architecture.

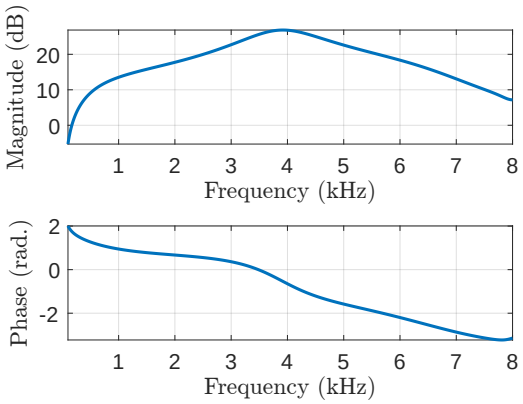

Figure 2: Magnitude and phase response of filter used to simulate the listener's hearing threshold.

Although Generalized Cross-Correlation (GCC) and similar methods can be applied to any setup with two or more microphones, some recent research has focused on localization models specifically designed for binaural systems [5, 6]. Recent efforts have integrated azimuth-dependent models of ITD and ILD, demonstrating that jointly considering both cues enhances azimuth estimation compared to using ITD alone [5–7]. However, these models often require prior training or calibration with the binaural input due to the significant variability in the frequency-dependent patterns of ITDs and ILDs across individuals, which can lead to performance degradation in different binaural setups. Methods also differ in how they integrate interaural information across time and frequency, with these variations largely reflecting different assumptions about source activity and interaction. In [5], authors proposed a framework that determines the likelihood of each source location based on a Gaussian Mixture Model (GMM) classifier, which learns the azimuth-dependent distribution of ITDs and ILDs from joint analysis of both binaural cues. However, many binaural localization methods have focused on scenarios with minimal reverberation or background noise. One approach to improving localization in more complex environments involves using model-based information about the spectral characteristics of sound sources in the acoustic scene to selectively weight binaural cues. This involves estimating models for both target and background sources during a training stage, using spectral features derived from isolated source signals [6]. In [8], an end-to-end binaural localization algorithm that estimates the azimuth using Convolutional Neural Network (CNN)s to extract features from the binaural signal was introduced.

Human auditory cognition includes complex neurological processes for localization. Although ILDs and ITDs are widely accepted to be the primary interaural cues that influence human sound source localization [1], there is no standardized way to characterise them. Precedence effect, spectral cues, head movement and other psychoacoustical processes affect sound localization in humans. There is no universally accepted method of measuring the correlation between human sound localization and the frequency-varying interaural cues. In [9, 10], to demonstrate the preservation of spatial cues, the error in interaural cues of the enhanced speech was computed using an Ideal Binary Mask (IBM) that selects the speech-active regions in the signal.

A relevant approach to measuring the accuracy with which spatial information is preserved and the subsequent accuracy of localization of speech sources in noisy and enhanced speech signals would be to employ a model that predicts the localization of the speech-in-noise in a manner highly correlated to a human listener. This paper sets out to research methods of Direction of Arrival (DOA) estimation that are not necessarily the best-performing but specifically follow the performance of the human listener in terms of binaural localization. The paper will focus on an end-to-end binaural localization model for speech in noisy and reverberant conditions, introducing

a lightweight Convolutional Recurrrent Network (CRN) that utilizes input features based on GCC-PHAT, which is a first step towards this goal. The model adds synthetic internal ear noise to an audio signal to simulate the effects of the frequency-dependent hearing threshold of a normal listener. The model is trained on binaural speech data to directly predict the source azimuth without limiting the localization to a predetermined azimuth-dependent distribution of interaural cues. The approach is evaluated using a listening test that was conducted using 15 normal hearing listeners, in which the participants were tasked to localize a target speaker in simulated noisy and reverberant conditions.

2. SYSTEM DESCRIPTION

2.1 Signal model

A binaural system is comprises a left and a right channel. The time-domain signal y_L received by the left channel is modeled as

$$y_L(n) = s_L(n) + v_L(n), \tag{1}$$

where s_L is the anechoic clean speech signal, v_L is the noise and n is the discrete-time index. The in-ear noise added signal $\tilde{y_L}$ is given by

$$\bar{y}_L(n) = h_e(n) * y_L(n) + e_L(n) \tag{2}$$

where $h_e(n)$ is the impulse response of the filter depicted in Fig. 2 and $e_L(n)$ is the white noise added to the filtered noisy signal. The right channel is described similarly with a R subscript. The model adds fictitious internal ear noise to an audio signal to simulate the effects of the frequency-dependent hearing threshold of a normal listener, assuming that the input speech in the stronger channel is at the normal level defined in [11] to be 62.35 dB SPL". The noise spectrum is taken from [11, 12] and, at a particular frequency, equals the pure-tone hearing threshold minus $10\log_{10}(C)$ where C is the critical ratio. The critical ratio, C, is the power of a pure tone divided by the power spectral density of a white noise that masks it; this ratio is approximately independent of level. Hearing loss can also be taken into account here by modifying the filter that reduces the signal level by the hearing loss at each frequency. To avoid having to add very high noise levels at low and high frequencies, it instead filters the input signal by the inverse of the desired noise spectrum and then adds white noise with 0 dB power spectral density. Figure 1 shows the block diagram of the proposed system. The raw time-domain signal is filtered with the in-ear frequency response shown in Fig. 2. The online implementation (`v_earnoise.m` Matlab function) of the ear-noise filter can be found in [13]. The in-ear noise-added signal is then used as the input to the neural network, which determines the target azimuth in the frontal azimuthal plane.

2.2 Localization network

2.2.1 *Input Feature Set*

The input feature of the proposed network consists of the GCC-PHAT for the pair of microphone signal frames $(\bar{\mathbf{y}}_\mathbf{L}, \bar{\mathbf{y}}_\mathbf{R})$, defined as

$$\mathbf{g}_{LR} = \text{IDFT}\left(\frac{\bar{\mathbf{Y}}_L}{|\bar{\mathbf{Y}}_L|} \odot \frac{\bar{\mathbf{Y}}_R^*}{|\bar{\mathbf{Y}}_R|}\right), \tag{3}$$

the Inverse Discrete Fourier Transform (IDFT) of the element-wise product of the normalized frequency-domain frames $\mathbf{Y}_L$ and $\mathbf{Y}_R$, where $\bar{\mathbf{Y}} = \text{DFT}(\bar{\mathbf{y}})$ and $|\mathbf{Y}|$ is the element-wise magnitude.

2.2.2 *Network architecture*

As shown in Fig. 1, the network is composed of a set of convolutional blocks, followed by an operation of flattening of the frequency and channel dimension. The resulting tensor is then used as input for a Gated Recurrent Unit (GRU) Recurrent Neural Network (RNN). Finally, a linear layer is applied to produce a 2-D output vector, $\hat{\mathbf{v}}$, representing the direction of the source's azimuth.

2.3 Loss function

The proposed model is trained using a modification of the cosine similarity given by

$$\mathcal{L}(\mathbf{v}, \hat{\mathbf{v}}) = 1 - \| \frac{\mathbf{v} \cdot \hat{\mathbf{v}}}{|\mathbf{v}||\hat{\mathbf{v}}|} \| \tag{4}$$

between the true and estimated directions $\mathbf{v}$ and $\hat{\mathbf{v}}$. The loss function (4) was designed so that the absolute value of the cosine similarity between the vectors is minimized, therefore not penalizing the effects caused by the front-back ambiguity, which are expected when employing only two microphones.

3. EXPERIMENTS

3.1 Dataset

To generate binaural speech data, monaural clean speech signals were obtained from the CSTR VCTK corpus [14]

and spatialized using the measured Binaural Room Impulse Responses (BRIRs) from [15] for training. The VCTK corpus contains approximately 13 hours of speech data from 110 English speakers with various accents. These recordings were used to create 2 s speech utterances, which were spatialized to produce left and right ear channels. The resulting dataset comprised 20,000 speech utterances, which were divided into training (70%), validation (15%), and testing (15%) sets. Diffuse isotropic speech-shaped noise was generated using uncorrelated noise sources uniformly distributed every 5° in the azimuthal plane [16], utilizing BRIRs from [15] which were recorded in a listening room with a T_{60} of 460 ms. The binaural signals were generated with the target speech positioned at a random azimuth in the frontal plane (-90° to $+90^\circ$) with the source positioned at a distance of 100 cm. For the training process, isotropic noise was added so that the average in dB of $(SNR_L,\ SNR_R)$, ranged between -25 dB and 25 dB. The evaluation set comprised speech signals spatialized with BRIRs from [17] with random target azimuths and isotropic noise added at random SNRs between -25 dB and 25 dB. The speaker was positioned at a 0° elevation and at a distance of 3 m. This ensured that training and evaluation sets contained binaural signals generated using different BRIRs to verify that the network generalised to different heads.

3.2 Training Setup

The 2 s input signals were sampled at 16 kHz, and a window size of 512 was used to generate the signal frames with a 75% overlap for a hop size of 25 ms. The parameters for the localization network are detailed in Fig. 1, which includes the tensor output shapes for each layer of the network. Convolutional layers employed a kernal size of (3, 3) throughout. Max pooling with a kernel size of 2 was applied to all convolutional layers except the last one. The Parametric Rectified Linear Unit (PReLU) activation function was utilized in all layers of the network, except for the RNN and Multi-Layer Perceptrons (MLP) output layers, which used hyperbolic tangent (tanh) activation, and the output layer, which employed sigmoidal activation. This architecture was taken from [18] and modified to work for binaural signals. The network has 850K parameters and is implemented using the PyTorch library, and the Adam optimizer was used for backpropagation. The network was trained for 80 epochs. The code for implementation is available online [1].

[1] https://github.com/VikasTokala/BiL

Method	Localization Error
SRP-PHAT	10.2°
WaveLoc-GTF [8]	3.0°
WaveLoc-CONV [8]	2.3°
BIL	1.2°

Table 1: Localization error compared to WaveLoc [8] methods.

3.3 Listening Tests

In the listening tests, 15 participants with normal hearing were tasked with localizing a target speaker within the frontal azimuthal plane. Using Beyerdynamic DT1990 Pro open-back headphones, the audio signals were delivered in a soundproof booth through an RME Fireface UCX II audio interface. The participants were required to listen to the noisy speech utterances and select the perceived azimuth using a MATLAB-based GUI. The azimuths were quantized at 15° intervals. Each participant listened to 36 speech utterances, which were evenly distributed across different SNRs and randomly assigned azimuths in the frontal azimuth plane. Three conditions of input SNR (iSNR) were used in the test: -15, 0 and +15 dB iSNR corresponding to "very noisy", "noisy" and "low noise" conditions, respectively.

4. RESULTS AND DISCUSSION

The model was evaluated using 275 speech utterances for each noisy input SNR ranging from -25 dB to +25 dB in steps of 5 dB. The localization error for the proposed method, denoted as BIL, is shown in Fig. 3a for different iSNRs. The azimuth θ of the target speaker's DOA in the frontal azimuth plane is then estimated using the Steered Response Power with Phase Transform (SRP-PHAT) algorithm [19, 20] and used for comparison. In extremely noisy conditions, such as -25 dB, the proposed method achieves a localization error of approximately 15°. Under similar iSNR conditions, the localization error for Steered Response Power (SRP) is considerably higher, around 40°. As the iSNR improves, the localization error for the proposed method decreases to below 5°, eventually reaching just under 1° at 25 dB iSNR. In contrast, the SRP method maintains an error between 10° and 20° even at higher iSNRs. The reduced performance of SRP at higher iSNRs can be attributed to reverberation, which

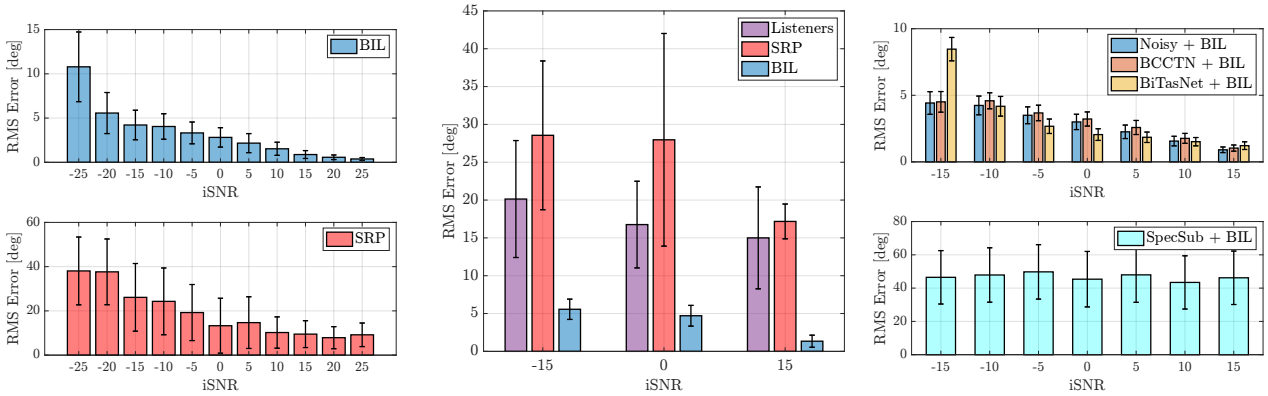

Figure 3: The plots show the localization error in noisy reverberant conditions (a) for the proposed method (BIL) and SRP, (b) for listeners compared with the proposed method and SRP and, (c) for signals processed by different enhancement methods evaluated by BIL.

causes multiple peaks in the correlation [18]. 1 shows the comparison of localization error with the WaveLoc methods proposed in [8]. These methods are also evaluated on BRIRs from [15] without the addition of external noise, and the values shown are taken from [8]. For similar conditions, the proposed method has lower error and outperforms both versions of the WaveLoc methods.

Figure 3b shows the localization error of human listeners compared with the proposed method and SRP for the three conditions of noisy signals as described in Sec. 3.3. The proposed method has a significantly lower localization error for all the iSNR conditions. Listeners had an average error of 20° in the very noisy condition of -15 dB and an average error of 15° in the low noise condition of 15 dB, given that there was no head movement to assist them. SRP-based localization had the highest localization error and standard deviation for the test samples. Previous studies have shown that human localization of speech and tones can have a localization error of up to 40° when noise and reverberation are present [1–3]. If the signals processed by enhancement methods produce a low localization error with the proposed method, it is very likely that the interaural cues of the signal are preserved, and human listeners will still localize the target speaker in the same azimuth as the original noisy signal.

Figure 3c demonstrates how the proposed method can be used to assess the performance of binaural speech enhancement methods in preserving the interaural cues and the spatial information of the target speaker. While there are well-known objective measures to evaluate noise reduction, speech intelligibility and quality, there are no standardised measures to assess the preservation of binaural cues after they are processed by enhancement algorithms. The upper plot in Fig. 3c shows the localization error for noisy signals at the iSNRs from -15 dB to 15 dB and the signals processed by Binaural Complex Convolutional Transformer Network (BCCTN) [9] and Binaural TasNet (BiTasNet) [21] at the same iSNRs. The binaural enhancement algorithms are designed to preserve the interaural cues in the noisy signal while enhancement, and they show a low localization error. At -15 dB, the BiTasNet shows a higher error compared to the noisy input signal, which indicates disruption in the interaural cues, and this is expected as the method was not designed to perform enhancement at -15 dB. As the iSNR improves, all the binaural enhancement methods show localization error under 5°, which signifies the preservation of interaural cues. From Fig 3a - Fig. 3c, it is evident that the proposed model has a monotonic relationship to SNR, i.e., the localization error decreases with increasing iSNR. Furthermore, other studies, including [1–3], show that human localization capability is monotonically proportional to SNR. Hence, the proposed method has been seen to be, as desired, highly correlated with human binaural localization - a conclusion which is supported by the subjective listening tests conducted. The lower plot in Fig. 3c shows the localization error obtained when the noisy signals are processed with bilateral spectral subtrac-

tion (SpecSub) [22], where no attempt is made at preserving binaural cues. The localization error is obtained around $45°$ as the testset contains signals which have azimuths distributed randomly between $\pm 90°$. If the binaural enhancement methods are being used for purposes other than human listening, the addition of in-ear noise can be omitted before performing localization.

5. CONCLUSION

This paper presented an end-to-end binaural localization model for speech in noisy and reverberant conditions. A CRN network utilizing GCC-PHAT features was introduced, and a listening test with 15 normal-hearing listeners showed that the model closely aligns with human perception, albeit with lower localization error. The model effectively evaluates the localization error of binaural speech enhancement algorithms, correlating with spatial information preservation and interaural cue retention. The key objective was to develop a DOA estimation method that mirrors human binaural localization rather than purely optimizing accuracy. The proposed method demonstrated significantly lower localization errors across all iSNR conditions. Listeners had average errors of $20°$ at -15 dB and $15°$ at 15 dB without head movement. SRP-based localization showed the highest error and variability and as iSNR improves, all binaural enhancement methods exhibit localization errors below $5°$, confirming interaural cue preservation. The model's localization error follows a monotonic relationship with SNR, aligning with human performance trends.

6. ACKNOWLEDGMENTS

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956369 and the UK Engineering and Physical Sciences Research Council [grant number EP/S035842/1].

7. REFERENCES

[1] M. D. Good and R. H. Gilkey, "Sound localization in noise: The effect of signal-to-noise ratio," *J Acoust Soc Am*, vol. 99, pp. 1108–1117, Feb. 1996.

[2] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in normal-hearing listeners," *J Acoust Soc Am*, vol. 105, pp. 1810–1820, Mar. 1999.

[3] M. L. Folkerts, E. M. Picou, and G. C. Stecker, "Spectral weighting functions for localization of complex sound. II. The effect of competing noise," *J Acoust Soc Am*, vol. 154, pp. 494–501, July 2023.

[4] N. Kopčo, V. Best, and S. Carlile, "Speech localization in a multitalker mixture," *J Acoust Soc Am*, vol. 127, pp. 1450–1457, Mar. 2010.

[5] T. May, S. van de Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, pp. 1–13, Jan. 2011.

[6] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, pp. 2122–2131, Nov. 2018.

[7] J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, pp. 1503–1512, July 2012.

[8] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end Binaural Sound Localisation from the Raw Waveform," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 451–455, May 2019.

[9] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural Speech Enhancement using Deep Complex Convolutional Recurrent Networks," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, (USA), 2023.

[10] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural Speech Enhancement using Deep Complex Convolutional Transformer Networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, (Seoul, South Korea), 2024.

[11] ANSI, "Methods for the calculation of the speech intelligibility index," ANSI Standard S3.5-1997 (R2007), American National Standards Institute (ANSI), 1997.

[12] C. V. Pavlovic, "Derivation of primary parameters and procedures for use in speech intelligibility predictions," *J Acoust Soc Am*, vol. 82, pp. 413–422, Aug. 1987.

[13] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997.

[14] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[15] J. Francombe, "IoSR Listening Room Multichannel BRIR Dataset - University of Surrey," 2017.

[16] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, (Tokyo, Japan), pp. 461–465, Sept. 2018.

[17] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-Ear head-related and binaural room impulse responses," *EURASIP J. on Advances in Signal Process.*, vol. 2009, p. 298605, July 2009.

[18] E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP Method for Universal Robust Multi-Source Tracking," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 19–28, 2024.

[19] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), Digital Signal Processing, pp. 157–180, Berlin Heidelberg: Springer-Verlag, 2001.

[20] E. Grinstein, E. Tengan, B. Çakmak, T. Dietzen, L. Nunes, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Steered Response Power for Sound Source Localization: A Tutorial Review," *EURASIP J. on Audio, Speech, and Music Process.*, vol. submitted, May 2024.

[21] C. Han, Y. Luo, and N. Mesgarani, "Real-Time Binaural Speech Separation with Preserved Spatial Cues," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 6404–6408, May 2020.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.