

BINAURAL TARGET SPEAKER EXTRACTION USING INDIVIDUALIZED HRTF

Yoav Ellinson

Faculty of Engineering, Bar-Ilan University, Israel
yoav.ellinson@biu.ac.il; 0009-0001-6116-2869

Sharon Gannot

Faculty of Engineering Bar-Ilan University, Israel
sharon.gannot@biu.ac.il; 0000-0002-2885-170X

Abstract—In this work, we address the problem of binaural target-speaker extraction in the presence of multiple simultaneous talkers. We propose a novel approach that leverages the individual listener’s Head-Related Transfer Function (HRTF) to isolate the target speaker. The proposed method is speaker-independent, as it does not rely on speaker embeddings. We employ a fully complex-valued neural network that operates directly on the complex-valued Short-Time Fourier transform (STFT) of the mixed audio signals, and compare it to a Real-Imaginary (RI)-based neural network, demonstrating the advantages of the former. We first evaluate the method in an anechoic, noise-free scenario, achieving excellent extraction performance while preserving the binaural cues of the target signal. We then extend the evaluation to reverberant conditions. Our method proves robust, maintaining speech clarity and source directionality while simultaneously reducing reverberation. A comparative analysis with existing binaural Target Speaker Extraction (TSE) methods shows that the proposed approach achieves performance comparable to state-of-the-art techniques in terms of noise reduction and perceptual quality, while providing a clear advantage in preserving binaural cues. Demo-page: <https://bi-ctse-hrtf.github.io>

Index Terms—Binaural target speaker extraction; Head-related transfer function; Complex-valued Neural network.

I. INTRODUCTION

Humans have a remarkable ability to focus on the voice of a single speaker, even in complex auditory environments with multiple speakers. Such a scenario is often referred to as the cocktail party problem [1]. While this ability may pose a challenge even to those with normal hearing, it presents a far greater obstacle for individuals with hearing impairments. Binaural Target Speaker Extraction (Bi-TSE) aims to selectively extract the desired speaker from multiple competing speakers in a binaural setting. In this context, preserving the binaural cues is critical: hearing aids and their associated algorithms must maintain the listener’s spatial awareness to ensure natural sound perception [2], [3]. Most hearing aids employ either a single microphone or multiple microphones per earpiece and can be broadly categorized into two groups: 1) Bilateral systems, where each device operates independently and provides a single-channel output per earpiece; 2) Shared systems, where all microphones across both earpieces are jointly utilized, yielding a dual-channel output.

Two important spatial cues are the Interaural Level Difference (ILD) and Interaural Time Difference (ITD), which capture magnitude and arrival-time differences between the left and right ears. In bilateral systems, in which each device processes its own channel independently, it is particularly challenging to preserve binaural cues. In contrast, Multi input Multi output (MIMO) solutions jointly process both channels, yielding more natural and spatially accurate outputs. For example, [4] presents a Binaural Linearly Constrained Minimum Variance (BLCMV) beamformer that separates sources while maintaining binaural cues. More recent work, such as [5], employs a Deep Neural Network (DNN) to process all microphone signals and produce left- and right-channel outputs, effectively preserving spatial information.

Unlike Blind Source Separation (BSS) algorithms, which separate all sources, TSE algorithms leverage prior knowledge on the target speaker(s) to isolate their voice. This prior knowledge—commonly referred to as a “clue” [6]—can take various forms. A widely used clue is a pre-recorded speech sample, often referred to as an enrollment utterance, which provides speaker-specific features [7]–[10]. Clues may also come from other modalities, such as visual data. For example, [11] uses a face image to extract the corresponding speech, while [12] combines audio and visual clues, with the audio clue being an enrollment utterance and the visual clue a video of the speaker speaking within the mixture.

While examining prior works, such as [13], we observed a particular challenge in separating two voices with highly similar characteristics, for example, speakers of the same gender. Consequently, we opted for an approach that does not rely on speaker-embedding-based clues. Extending beyond a single microphone enables us to exploit spatial information and more effectively isolate the desired speaker. For instance, [14] uses either the Relative Transfer Function (RTF) or the Direction of Arrival (DOA) of the target speaker relative to a microphone array, while [15], [16] rely on the DOA of the desired speaker and use it as a spatial clue. These spatially guided methods consistently outperform single-microphone solutions.

In this work, we propose Binaural Target Speaker Extraction using HRTFs (Bi-TSE-HRTF), a target speaker extraction method that exploits the individual listener’s HRTF, aligned with the target speaker’s DOA, as the primary extraction clue. Our approach employs a complex-valued DNN that operates

directly in the STFT domain. The proposed architecture produces binaural audio outputs, ensuring that spatial cues are fully preserved. Moreover, in reverberant environments, the model isolates the direct path's HRTF, thereby dereverberating the signal and hence further enhancing its binaural characteristics. The method is thoroughly assessed using simulated Room Impulse Responses (RIRs) and standard HRTF databases, and compared with two competing methods, a beamformer-based solution and a DNN-based solution.

II. PROBLEM FORMULATION

We address the problem of multiple concurrent speakers captured by two microphones mounted on the listener's ears, assuming a noiseless environment and a fixed source–listener radius. For clarity, we focus on the case of two concurrent speakers, as formulated below.

Denote the desired speaker's DOA by (θ_d, ϕ_d) and the interferer's by (θ_i, ϕ_i) . Let $\mathbf{x}_b(n, k)$ denote the STFT of the binaural signal, with $n \in [0, N-1]$ and $k \in [0, K-1]$ the time-frame and frequency-bin indices, respectively. The binaural signal is therefore formulated as:

$$\mathbf{x}_b(n, k) = s_d(n, k) \mathbf{h}(\theta_d, \phi_d, k) + s_i(n, k) \mathbf{h}(\theta_i, \phi_i, k) \quad (1)$$

where $s_d(n, k)$ and $s_i(n, k)$ are the anechoic speech signals of the desired speaker and the interfering speaker, respectively. Let $\gamma \in \{d, i\}$ denote the speaker index, where d refers to the desired source and i to the interferer. The acoustic transfer from source γ to both earpieces is described by $\mathbf{h}(\theta_\gamma, \phi_\gamma, k)$, which models the Transfer Function (TrF) as a superposition of HRTFs corresponding to the DOAs of all acoustic reflections in the enclosure, with the HRTF, $\mathbf{h}_{\text{hrtf}}(\theta, \phi, k)$, characterizing the listener's response to a wave arriving from DOA (θ, ϕ) .

According to the precedence effect [17], the direct-path HRTF dictates the perceived source DOA. Thus, although $\mathbf{h}(\theta_\gamma, \phi_\gamma, k)$ encompasses all reflections, it is indexed by the direct-path DOA for clarity. Finally, define $\bar{\mathbf{h}}(\theta_\gamma, \phi_\gamma, t)$, with t the filter's tap index, as the time-domain representation of $\mathbf{h}(\theta_\gamma, \phi_\gamma, k)$, and similarly define the Head-Related Impulse Response (HRIR) at the DOA θ, ϕ as $\bar{\mathbf{h}}_{\text{hrir}}(\theta, \phi, t)$. Finally, the target signal, $\tilde{\mathbf{s}}_d = [\tilde{s}_d^l, \tilde{s}_d^r]^\top$, is a concatenation of the desired speech signal as received at the left and right ears, respectively.

Given the individual HRTF of the hearing-aid wearer corresponding to the desired source DOA (θ_d, ϕ_d) , denoted by the binaural response $\mathbf{h}_{\text{hrtf}}(\theta_d, \phi_d, k)$, our method aims to extract $\tilde{\mathbf{s}}_d$ from the mixed signal $\mathbf{x}_b$. The extracted signal is denoted by $\hat{\tilde{\mathbf{s}}}_d$. It is required that the binaural output satisfy the following conditions: 1) preservation of the spatial cues of the desired source, and 2) dereverberation of the desired speech output by retaining only the direct-path HRTF. In this work, we focus on *individualized* HRTF database, namely the set of HRTFs of the hearing aid wearer across all azimuth and elevation directions, which should be given to the algorithm. HRTF personalization using anthropometric measurements [18].

III. PROPOSED METHOD

We propose a DNN-based model to extract a desired speech signal corresponding to a given HRTF clue.

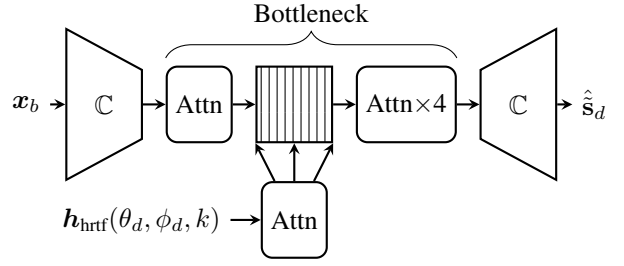


Fig. 1: A block diagram of the proposed method, where $\mathbf{h}_{\text{hrtf}}(\theta_d, \phi_d, k)$ denotes the HRTF of the desired speaker's DOA, $\mathbf{x}_b$ represents the mixed binaural signal, and $\hat{\tilde{\mathbf{s}}}_d$ represents the estimated desired signal, both in the STFT domain.

A. Architecture

The proposed model architecture is built upon the approach presented in [14], with modifications to handle dual-channel, complex-valued inputs rather than treating the real and imaginary components independently. The proposed model employs a U-Net architecture, incorporating the HRTF at the bottleneck. The individual modules comprising the architecture are described in detail below and depicted in Fig. 1.

Encoder–Decoder Blocks of the U-Net: The encoder consists of five down-convolution blocks, each comprising a 2D convolutional layer, global layer normalization, and a complex-valued ReLU activation defined as

$$\mathbb{C}\text{ReLU}(z) = \text{ReLU}(\mathcal{R}(z)) + j \text{ReLU}(\mathcal{I}(z)), \quad (2)$$

where $\text{ReLU}(x) = \max(0, x)$, $\mathcal{R}(z)$ and $\mathcal{I}(z)$ denote the real and imaginary parts of z , and j is the imaginary unit. The decoder mirrors the encoder with five up-convolution blocks based on transposed convolutions. Skip connections are employed between corresponding encoder and decoder blocks to improve reconstruction accuracy. The decoder further includes a final convolutional layer without an activation function, allowing the output to represent both positive and negative real and imaginary values, which is required for the complex STFT representation.

Bottleneck: The extraction clue is applied at the bottleneck of the encoded input features, with self-attention layers used both before and after its application. Unlike conventional approaches, these self-attention layers operate directly on complex-valued tensors by replacing the softmax activation with the complex-valued $\mathbb{C}\text{ReLU}$. To capture cross-channel (left–right) information in the HRTF, a self-attention mechanism is employed, enabling the channels to attend to one another via learned weights. The resulting matrix is flattened and projected via a fully-connected layer to match the dimensionality of the encoded mixture, yielding the HRTF embedding. Similarly, self-attention is applied to the encoded mixed signal. The replicated HRTF embedding is then combined with each time frame of the encoded mixture through element-wise multiplication, thereby imposing the extraction clue. The resulting tensor is processed by four additional self-attention layers and a fully connected layer before decoding.

B. Loss function

The model was initially trained using only the Scale Invariant Signal to Distortion Ratio (SI-SDR) loss (3).

$$\text{SI-SDR}(s, \hat{s}) = 20 \log_{10} \left(\frac{\|\beta s\|}{\|\beta s - \hat{s}\|} \right); \quad \beta = \frac{\hat{s}^\top s}{\|s\|^2}, \quad (3)$$

where s and $\hat{s}$ are vectors comprising all time-domain samples of the target and estimated signals, respectively. However, relying solely on the SI-SDR loss proved challenging in preserving accurate spectral details, resulting in noticeable musical noise artifacts in the extracted signals. Musical noise refers to tonal distortions that occur due to inaccuracies or rapid variations in the estimated signal's frequency-domain representation [19]. Such artifacts are commonly observed in signal enhancement tasks when spectral continuity and consistency are not adequately enforced. To mitigate these artifacts, we introduced an additional frequency-domain loss, the L_1 -Mean Absolute Error (MAE) loss:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{2KN} \sum_{c=\{l,r\}} \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \left| |\tilde{s}_d^c(n, k)| - |\hat{s}_d^c(n, k)| \right|. \quad (4)$$

Here, $c = \{l, r\}$ denotes the output channels, left or right. To improve data utilization, both speakers were extracted from each mixed signal in each training step by alternating the HRTF clues and averaging the SI-SDR losses. The overall training loss is therefore defined in (5), as a weighted combination of the SI-SDR and L_1 -MAE losses:

$$\mathcal{L} = \frac{1}{2} \sum_{\gamma=\{d,i\}} \left(-\text{SI-SDR}(\tilde{s}_\gamma, \hat{s}_\gamma) + \alpha \cdot \mathcal{L}_{\text{MAE}}(\tilde{s}_\gamma, \hat{s}_\gamma) \right). \quad (5)$$

Due to the complexity of multi-objective optimization in high-dimensional space, we adopted a multi-stage training strategy, in which we first train the algorithm using only the SI-SDR loss (i.e., $\alpha = 0$). Once a sufficient performance level has been reached, the L_1 -MAE loss is introduced (namely, $\alpha > 0$), improving both spectral accuracy and audio quality. The final stage fine-tunes the model using only the SI-SDR loss, as the L_1 -MAE loss is prone to overfitting to the training data.

IV. EXPERIMENTAL STUDY

In this section, we describe the data-generation process and experimental setup, evaluate the performance of the proposed approach, and compare it with previous TSE methods.

A. Data Generation

Binaural Impulse Responses: The binaural Impulse Responses (IRs) are formed by superimposing attenuated and time-shifted HRIRs, corresponding to acoustic reflections generated via the image method [20], as expressed in (6).

$$\bar{h}(\theta_\gamma, \phi_\gamma, t) = \sum_{\ell=0}^L a_\ell \cdot \bar{h}_{\text{hrr}}(\theta_\ell, \phi_\ell, t - \tau_\ell) \quad (6)$$

where L is the reflection order. The direct path ($\ell = 0$) corresponds to the HRIR of the desired or interfering source at the listener's head for the DOAs $(\theta_\gamma, \phi_\gamma)$, and the gain a_ℓ

and the time delay τ_ℓ , obtained from the image method, are applied by scaling and time-shifting each HRTF component.

The RIEC HRTF Dataset [21], [22]: We utilized the RIEC database, which comprises measurements from 105 subjects, each with 865 directions. For this study, we focused on a single subject. The Spatially Oriented Format for Acoustics (SOFA) HRTF Toolbox [23], [24] facilitates easy retrieval of the appropriate HRTF for any desired DOA. We employed the SofaMyRoom framework [25] to simulate both anechoic and reverberant conditions, following the formulation in (6). This framework positions a listener, characterized by their measured HRTF, inside a shoebox-shaped room with configurable wall materials to achieve a specified reverberation time. It also supports zero-order simulations for anechoic scenarios. The reverberation time was sampled as $T_{60} \sim \mathcal{U}[0.2, 0.8]$, and the listener position is drawn uniformly within the room. The overall simulation setup is illustrated in Fig. 2.

Speech Datasets We trained and validated our model on the WSJ0-CSR1 dataset [26].

B. Algorithm Setup

The speech signals, sampled at 16 kHz, are randomly selected, cropped, or zero-padded to 5 s, then transformed with a 512-sample STFT (75% overlap). The model is trained in a supervised manner: in the anechoic setting, targets are generated by convolving the desired speech with the corresponding HRTF; in the reverberant setting, by simulating a room and keeping only the first arrival (zero-order reflection). The HRTFs, used as the extraction clue, are downsampled from 48 kHz to 16 kHz and converted to the frequency domain via Fast Fourier Transform (FFT).

C. Metrics

To evaluate our method, we use SI-SDR [28], wide- and narrow-band Perceptual Evaluation of Speech Quality (PESQ) [29], and Short-time Objective Intelligibility (STOI) [30] as intrusive metrics of speech quality and intelligibility. In addition, we employ DNS-MOS [31], a non-intrusive DNN-based Mean Opinion Score (MOS) predictor estimating speech quality (SIG), background noise quality (BAK), and overall quality (OVRL). Furthermore, to assess binaural cue preservation, we compare the ILD-ITD p.d.f.s of the extracted and target signals, where close agreement indicates preserved spatial cues. These metrics are computed using the procedure in [27].

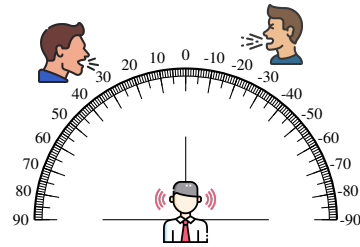


Fig. 2: A simulation illustration: two concurrent speakers: the desired speaker at $\theta_d = 40^\circ$ (left) and the interferer at $\theta_i = -30^\circ$ (right), both at elevation ϕ . Images source: freepik.com

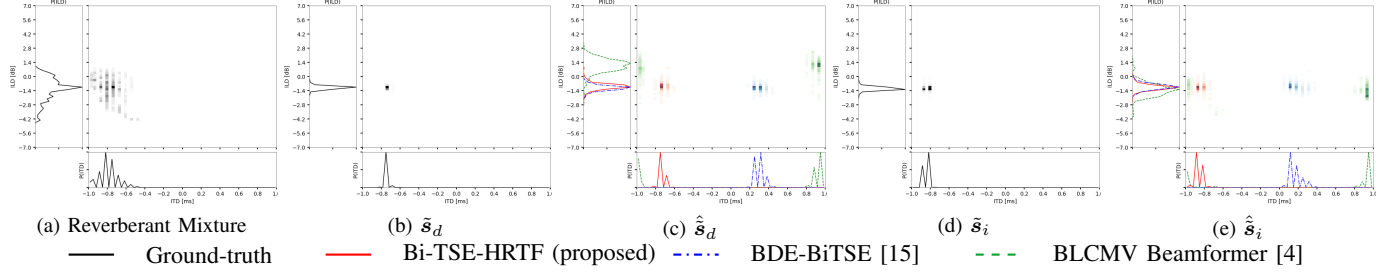


Fig. 3: The joint Probability Density Functions (p.d.f.s) of ITD [ms] and ILD [dB] for $\theta_d = -60^\circ$, $\theta_i = -90^\circ$, and $\phi_{d,i} = -10^\circ$, with $T_{60} = 0.63$ s, for the frequency band centered at 500 Hz, where the ILD is less pronounced. $\tilde{s}_d$ and $\tilde{s}_i$ denotes the anechoic target signals. Graphs produced by [27].

TABLE I: Each evaluation dataset contains 1500 speech mixtures, with both speakers extracted, yielding 3000 samples. Mixtures were created with Signal-to-Interference Ratio (SIR) drawn from $\mathcal{U}[-5, 5]$ dB, and 100% overlap. Best scores are highlighted in boldface.

Dataset	Method	DNSMOS (OVRL\SIG\BAK)	SI-SDR (dB) $\uparrow$	WB PESQ $\uparrow$	NB PESQ $\uparrow$	STOI $\uparrow$	Δ ILD [dB] $\downarrow$	Δ ITD [ms] $\downarrow$
WSJ0	Unprocessed	(2.845\3.376\3.411)	0	1.25	1.57	0.72	-	-
	BLCMV Beamformer	(3.044\3.444\3.748)	-0.42	3.48	3.84	0.95	0.745	0.613
	BDE-BiTSE	(3.230\3.543\3.951)	25.59	3.77	4.01	1.0	0.126	0.484
	Bi-TSE-DOA	(2.965\3.376\3.660)	16.44	2.43	3.09	0.97	0.165	0.990
	Bi-TSE-HRTF-RI	(3.110\3.504\3.764)	21.39	2.93	3.57	0.99	0.085	0.0044
	Bi-TSE-HRTF (proposed)	(3.18\3.501\3.911)	21.41	3.47	3.88	0.99	0.0756	0.0035
WSJ0 (Reverb)	Unprocessed	(2.348\3.069\2.783)	-1.9	1.15	1.44	0.69	-	-
	BLCMV Beamformer	(1.840\2.612\2.038)	-1.11	1.35	1.88	0.74	0.430	0.593
	BDE-BiTSE	(3.141\3.438\3.952)	14.8	3.09	3.56	0.98	0.204	0.386
	Bi-TSE-DOA	(2.681\3.196\3.324)	8.29	1.82	2.44	0.91	0.296	0.990
	Bi-TSE-HRTF-RI	(3.030\3.362\3.838)	11.18	2.01	2.74	0.95	0.174	1.001
	Bi-TSE-HRTF (proposed)	(2.86\3.348\3.481)	10.67	2.17	2.79	0.94	0.186	0.0082

TABLE II: Comparison of Δ ITD, $\hat{\theta}$, and $\Delta\theta$ for $\theta_d = -60^\circ$.

Method	Δ ITD [ms]	$\hat{\theta}$ [$^\circ$]	$\Delta\theta$ [$^\circ$]
BLCMV Beamformer	1.7	> 90	> 150
BDE-BiTSE	1.0667	30	90
Bi-TSE-HRTF	0.0	-60	0

D. Competing Methods

To ensure a fair comparison, all baseline methods were evaluated under an identical setup, namely binaural TSE with only two microphones (one per ear). Accordingly, we compared our method against the BLCMV beamformer [4] and a recent DNN-based model [15], both specifically designed for this scenario. The competing DNN, referred to as BDE-BiTSE, was trained following the procedure described in the paper, using the exact same training data as ours. We ensured that the strongest variant reported in the paper (BDE+CDF+IPD+SDF) was used for comparison.

As an ablation study, we trained two variants of our model. The first, Bi-TSE-HRTF-RI, is a real-valued model in which the complex-valued input features are decomposed into their RI components and concatenated, yielding four input channels rather than two. The same decomposition is applied to the HRTF clue. The second, Bi-TSE-DOA, is a complex-valued model that incorporates the DOA as a clue rather than the HRTF. Here, the DOA is represented as a one-hot-encoded vector with the same resolution as the available HRTFs and passed through a fully connected layer to match the bottleneck's dimensionality. For both variants, the loss function (5) remains unaltered, with the target defined as the binaural rendering of the desired speech signal using the correct HRTF.

E. Results

In Table I, we present a comprehensive comparison of our method with competing algorithms in both anechoic and reverberant conditions. In terms of SI-SDR, PESQ, and DNS-MOS, our approach achieves competitive results in the anechoic setting, but underperforms in reverberant environments compared with BDE-BiTSE. Nevertheless, our model clearly preserves binaural cues, outperforming the baseline methods, as evidenced by the ITD and ILD measures. The importance of minimizing Δ ITD is evident in systems that do not explicitly constrain it. Since ITD reflects interaural arrival-time differences [3], inaccuracies lead to spatial errors and shifts in the perceived target location. To quantify this effect, we estimate the output DOA by computing ITD values for all candidate directions, forming a DOA-ITD lookup table, and comparing the estimated direction $\hat{\theta}$ to the ground truth. Owing to the computational cost, we report a single illustrative example in Table II. The results indicate that competing methods exhibit substantial localization errors, whereas the proposed approach preserves spatial consistency.

Figure 3 shows the joint ITD-ILD p.d.f.s for the extracted and target signals. The close alignment of the proposed method with the target distribution confirms effective preservation of binaural cues. In contrast, competing methods exhibit noticeable deviations, resulting in shifted perceived source locations. The increased concentration of the joint distribution further indicates a dereverberation effect, with the direct path dominating the output.

Additionally, we evaluate two model variants: Bi-TSE-

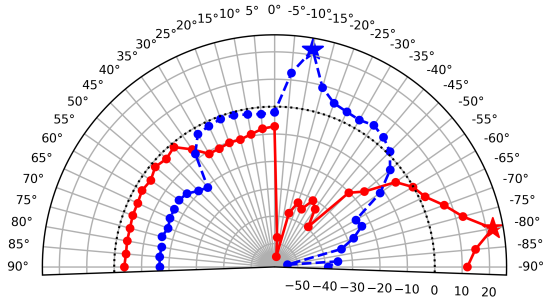


Fig. 4: A polar plot of SI-SDR [dB] versus DOA, with the target at $\theta_d = -75^\circ$ (red) and the interferer at $\theta_i = -10^\circ$ (dashed blue). Stars mark true DOAs;

DOA and Bi-TSE-HRTF-RI. The DOA-based model yields lower SI-SDR and PESQ, fails to suppress the interferer, and does not preserve binaural cues—especially under reverberation—highlighting the central role of HRTFs in our formulation. The RI model achieves SI-SDR and binaural cue preservation comparable to those of the complex-valued model in anechoic conditions, but yields lower PESQ scores and degraded cue preservation under reverberation. To demonstrate spatial resolution, we employ a “scanning” procedure in which signals are extracted using HRTFs over azimuths in $[-90^\circ, 90^\circ]$, and the corresponding SI-SDR is computed with respect to the target. Plotting these values yields a beam-pattern-like diagram (Fig. 4) with a peak at the true DOA, where the main-lobe width reflects the model’s stability to mild HRTF deviations.

V. CONCLUSIONS

This paper presents a speaker-independent TSE method that exploits the listener’s HRTFs in both anechoic and reverberant environments, providing a personalized solution. By conditioning the extraction process on individualized HRTFs, the proposed approach preserves binaural structure while maintaining speech quality. Comparative analyses with existing binaural TSE methods demonstrate substantial improvements in spatial consistency, while preserving high speech quality. These results highlight the critical role of the listener’s HRTF in anchoring spatial cues, enabling accurate and consistent target-speaker extraction across diverse acoustic conditions.

REFERENCES

- [1] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.
- [2] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*, MIT Press, 1997.
- [3] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, *Audio Source Separation and Speech Enhancement*, chapter Binaural Speech Processing with Application to Hearing Devices, Wiley, 2018.
- [4] E. Hadad, S. Doclo, and S. Gannot, “The binaural LCMV beamformer and its performance analysis,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 543–558, 2016.
- [5] C. Han, Y. Luo, and N. Mesgarani, “Real-time binaural speech separation with preserved spatial cues,” in *Proc. ICASSP*, 2020, pp. 6404–6408.
- [6] K. Žmolíková, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, “Neural target speech extraction: An overview,” *IEEE Signal Process. Mag.*, vol. 40, no. 3, pp. 8–29, 2023.
- [7] M. Delcroix, K. Žmolíková, K. Kinoshita, A. Ogawa, and T. Nakatani, “Single-channel target speaker extraction and recognition with speaker-beam,” in *Proc. ICASSP*, 2018, pp. 5554–5558.
- [8] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, “Deep extractor network for target speaker recovery from single-channel speech mixtures,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2018.
- [9] A. Eisenberg, S. Gannot, and S. E. Chazan, “Single-microphone speaker extraction using unified time-frequency siamese u-net,” in *Proc. EUSIPCO*, 2022, pp. 762–766.
- [10] C. Xu, W. Rao, E. S. Chng, and H. Li, “SpEx: Multi-scale time-domain speaker extraction network,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.
- [11] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, “FaceFilter: Audio-visual speech separation using still images,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2020.
- [12] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, “Multimodal speakerbeam: Single-channel target speech extraction with audio-visual speaker clues,” in *Proc. Interspeech*, 2019, pp. 2718–2722.
- [13] M. Delcroix, T. Ochiai, K. Žmolíková, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, “Improving speaker discrimination of target speech extraction with time-domain speakerbeam,” in *Proc. ICASSP*, 2020, pp. 691–695.
- [14] A. Eisenberg, S. Gannot, and S. E. Chazan, “End-to-end multi-microphone speaker extraction using relative transfer functions,” *arXiv:2502.06285*, 2025.
- [15] Y. Wang, J. Zhang, C. Jiang, W. Zhang, Z. Ye, and L. Dai, “Leveraging boolean directivity embedding for binaural target speaker extraction,” in *Proc. ICASSP*, 2025.
- [16] H. Meng, Q. Zhang, X. Zhang, V. Sethu, and E. Ambikairajah, “Binaural selective attention model for target speaker extraction,” in *Proc. Interspeech*, 2024, pp. 4323–4327.
- [17] H. Haas, “The influence of a single echo on the audibility of speech,” *Acustica*, vol. 1, pp. 49–58, 1951.
- [18] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, “HRTF personalization using anthropometric measurements,” in *Proc. WASPAA*, 2003, pp. 157–160.
- [19] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Signal Process. Lett.*, vol. 2, no. 2, pp. 345–349, 1994.
- [20] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
- [21] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, “Dataset of head-related transfer functions measured with a circular loudspeaker array,” *J. Acoust. Soc. Am.*, vol. 35, no. 3, pp. 159–165, 2014.
- [22] “RIEC HRTF database,” <http://www.riec.tohoku.ac.jp/pub/hrtf>.
- [23] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, and H. Ziegelwanger, “Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions,” in *AES Conv.*, 2013.
- [24] P. Majdak, F. Brinkmann, J. De Mynke, M. Mihocic, and M. Noisternig, “Spatially oriented format for acoustics 2.1: Introduction and recent advances,” *J. Acoust. Soc. Am.*, vol. 70, pp. 565–584, 2022.
- [25] R. Barumerli, D. Bianchi, M. Geronazzo, and F. Avanzini, “SofaMyRoom: A fast and multiplatform shoebox room simulator for BRIR dataset generation,” *arXiv:2106.12992*, 2021.
- [26] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) complete,” <https://doi.org/10.35111/ewkm-cg47>.
- [27] C. Faller and J. Merimaa, “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,” *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, 2004.
- [28] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR: Half-baked or well done?,” in *Proc. ICASSP*, 2019, pp. 626–630.
- [29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ),” in *Proc. ICASSP*, 2001, pp. 749–752.
- [30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [31] C. K. A. Reddy, V. Gopal, and R. Cutler, “DNSMOS: A non-intrusive perceptual objective speech quality metric,” in *Proc. ICASSP*, 2021, pp. 6493–6497.