DOI: 10.61782/fa.2025.0741

FORUM ACUSTICUM
EURONOISE 2025

RELATIVE TRANSFER MATRIX-BASED BINAURAL SIGNAL DENOISING OF HEAD-MOUNTED MICROPHONE ARRAY RECORDINGS

Manish Kumar **Amy Bastine** **Lachlan Birnie**
Sandra Arcos Holzinger **Prasanga N. Samarasinghe** **Thushara D. Abhayapala**
Audio & Acoustic Signal Processing Group
The Australian National University, Canberra, Australia

ABSTRACT

Head-mounted microphone arrays are increasingly prevalent in applications ranging from virtual reality to assistive hearing devices. Accurately enhancing binaural signals from these devices is crucial yet challenging in complex acoustic environments characterized by multiple sound sources and significant reverberation. The Relative Transfer Matrix (ReTM) approach, which generalizes relative transfer functions for multiple simultaneously active sources and receivers, has demonstrated success in speech denoising. This paper addresses the problem of binaural signal denoising by utilizing ReTM derived from head-mounted microphone array recordings. Our key contribution is adapting the ReTM computation to accommodate the user's head movements based on head-tracking data, which enhances the fidelity of the denoising process. We demonstrate this application with an augmented reality (AR) glass setup, equipped with four microphones on the frame and two over-ear microphones. The noise-only ReTM, computed between the on-frame and over-ear microphones across various head orientations, is employed to estimate and subsequently subtract noise from the binaural signal. The simulation results indicate that a higher resolution of ReTM-Dictionary leads to better speech quality (STOI, PESQ, SegSNR) scores, with improved preservation of binaural cues (ITD and ILD).

*__Corresponding author__: Manish.Kumar@anu.edu.au.

Keywords: *Relative Transfer Matrix, Binaural Processing, Speech Denoising, Augmented Reality, STFT*

1. INTRODUCTION

The binaural signal denoising algorithms has been an active area of research due to its wide range applications, including hearing aid technologies [1–3], cochlear implants [4,5], Augmented and Virtual Reality (AR/VR) [6,7], and hands-free communication [8,9]. The primary goal of the binaural denoising algorithm is to improve speech intelligibility and perceptual quality by preserving spatial auditory cues critical for source localization, accurate auditory scene perception, and immersive listening experiences [2,3].

The binaural denoising algorithms can be broadly classified into two categories: model-based and data-driven. The model-based approaches use classical signal processing techniques such as beamforming (MVDR, GSC, BLCMV, BMVDR) [10–13], spectral subtraction [14], Wiener filtering [15, 16], and Coherence-based filtering [3, 17]. These methods are computationally efficient, making them appropriate for real-time applications such as hearing aids or AR headsets. However, they often rely on explicit acoustic models and assumptions about the environment, making them struggle to generalize in complex or dynamic acoustic conditions [2, 18]. In contrast, the data-driven methods rely on machine learning techniques to learn mappings directly from data, making them more adaptable to diverse noise environments. These approaches often involve spatial feature extraction [19], self-supervised and contrastive learning [20], and

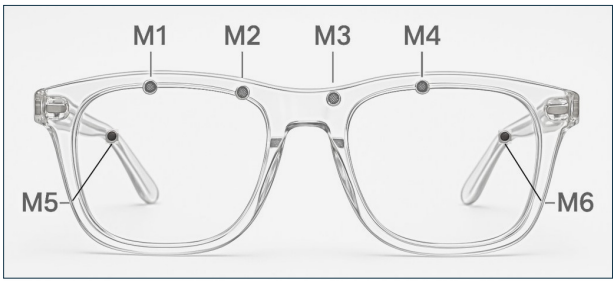

Figure 1: Layout of microphone array embedded in a pair of spectacles

convolutional neural networks (CNNs) [21, 22]. Although data-driven methods can achieve higher performance in challenging scenarios, they typically require large training datasets and may be computationally demanding [23].

In this paper, we present a model-based binaural denoising algorithm that leverages the Relative Transfer Matrix (ReTM) [24] as a spatial feature for AR or head-mounted microphone-array applications. Extending the theoretical framework established in [24, 25], we develop a ReTM dictionary that captures spatial transfer characteristics across different head orientations, an essential capability for dynamic AR environments. For binaural signal denoising, we select the appropriate ReTM from the dictionary and estimate noise characteristics at the binaural channels, followed by signal subtraction to get the denoised signal. The SMIR [26] based numerical simulation results validate the effectiveness of the proposed binaural denoising algorithm. The method does not require prior knowledge of the number of speech and noise sources, nor microphone locations, and can be extended to any configuration with more than three microphones. However, we assume that noise sources are continuously active, while speech sources are present intermittently.

The remainder of this paper is structured as follows: Section 2 introduces the system model and outlines the estimation of the Relative Transfer Matrix (ReTM) using a covariance-based approach. Section 3 details the construction of the ReTM dictionary and presents the proposed binaural denoising methodology. In Section 4, we evaluate the effectiveness of ReTM-based denoising through numerical simulations using the SMIR generator [26]. Finally, Section 5 summarizes the key findings and discusses potential directions for future research.

2. PROBLEM FORMULATION

In this section, we briefly describe the system model, followed by discussing the ReTM as a spatial feature, and how it can be estimated from the source signal.

2.1 System Model

Consider a reverberant environment in which a user wears a pair of spectacles equipped with K microphones, as illustrated in Fig.1). The signals captured by these microphones in the Short-Time Fourier Transform (STFT) domain are denoted by $M_k(f,t)$, where $k = \{1,\cdots,K\}$. Also, let there are $\mathcal{L}$ sound sources in the environment, comprising both speech and noise, represented as $S_\ell(f,t)$, $\ell = \{1,\cdots,\mathcal{L}\}$. To facilitate the ReTM estimation, we divide the microphones into two subgroups: a target group $\{A\}$ and a reference group $\{B\}$. The target group consists of K_A microphones, while the reference group consists of K_B microphones, with $K_A \leq K_B$. The signal components received by each microphone group can be represented in matrix form as follows,

$$\boldsymbol{M}_{\mathrm{A}}(f,t) = \boldsymbol{H}_{\mathrm{A}}(f)\boldsymbol{S}(f,t) \quad (1)$$

$$\boldsymbol{M}_{\mathrm{B}}(f,t) = \boldsymbol{H}_{\mathrm{B}}(f)\boldsymbol{S}(f,t), \quad (2)$$

where $\boldsymbol{M}_{\mathrm{A}}(f,t) = [M_1,\cdots,M_{K_A}]^T$, $\boldsymbol{S}(f,t) = [S_1,\cdots,S_{\mathcal{L}}]^T$, $[\cdot]^T$ denotes matrix transpose, and $\boldsymbol{H}_{\mathrm{A}} \in \mathbb{C}^{Q_A\times\mathcal{L}}$ is a matrix with elements defined by the acoustic transfer functions. Note, we have excluded microphone thermal noise from this formulation to focus on the core theoretical development presented in the following section.

2.2 Relative Transfer matrix (ReTM)

The ReTM characterizes the spatial mapping between two microphone groups in response to a given set of sound sources [24]. The ReTM ($\boldsymbol{\mathcal{R}}_{AB}(f)$), between microphone groups - $\{A\}$ (target) and $\{B\}$ (reference) in response to the sound sources can be modeled as [24]

$$\boldsymbol{M}_{\mathrm{A}}(f,t) = \boldsymbol{\mathcal{R}}_{\mathrm{AB}}(f)\boldsymbol{M}_{\mathrm{B}}(f,t). \quad (3)$$

The theoretical definition of the ReTM [24] is found by multiplying (2) from left by a suitable pseudo-inverse of $\boldsymbol{H}_{\mathrm{B}}$ and substituting for $\boldsymbol{S}$ in (1), resulting in

$$\boldsymbol{\mathcal{R}}_{\mathrm{AB}}(f) \triangleq \boldsymbol{H}_{\mathrm{A}}(f)\boldsymbol{H}_{\mathrm{B}}^{\dagger}(f), \quad (4)$$

where $(\cdot)^\dagger$ denotes Moore–Penrose inverse, assuming it is valid.

Note, the ReTM (4) is a matrix defined solely by the spatial characteristics of the sound sources, microphone receivers, and the surrounding acoustic environment [24].

2.3 ReTM Estimation

The ReTM estimation is based on covariance matrices [24, 25], a methodology widely used in the MUSIC source localization technique [27]. The auto and cross-covariance matrix of the received signals at microphone groups - $\{A\}$ and $\{B\}$ can be defined as,

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f) &\triangleq E\{\boldsymbol{M}_{\mathrm{A}}(f,t)\boldsymbol{M}_{\mathrm{A}}{}^{*}(f,t)\} \\ \boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f) &\triangleq E\{\boldsymbol{M}_{\mathrm{B}}(f,t)\boldsymbol{M}_{\mathrm{A}}{}^{*}(f,t)\} \end{aligned} \tag{5}$$

where $[\cdot]^*$ is conjugate transpose, and $E\{\cdot\}$ denotes the expectation, which can be found from averaged time frames. The ReTM is estimated by multiplying the pseudo-inverse of $\boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f)$ to $\boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f)$ [24, 25], as

$$\boldsymbol{\mathcal{R}}_{\mathrm{AB}}(f) = \boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f)\boldsymbol{\mathcal{P}}^{\dagger}_{\mathrm{BA}}(f). \tag{6}$$

Note that the estimated ReTM in (6) is an approximation only, as the microphones inevitably experience additive thermal noise in practice [24].

3. BINAURAL SIGNAL DENOISING WITH RETM-DICTIONARY

Consider a scenario where a user wearing a microphone array integrated into AR glasses wants to converse with a friend (the target speaker) in the same room. A loud static noise source, such as a television, interferes with the speech signal, making comprehension difficult. To denoise the speech signal in a multi-microphone setting with static noise, we previously presented a ReTM-based approach in [25]. However [25] assumes that the microphone array remains stationary with respect to the noise source. In real-world scenarios, such as a person wearing AR glasses with embedded microphones, this assumption often does not hold. During conversation, the user may move their head, nod, or make other gestures, resulting in changes to the array's orientation (azimuth and elevation). These changes to the array orientations make it challenging to apply a static ReTM for effective denoising, as the spatial relationship between the microphones and the noise source is no longer static.

To overcome this limitation, we propose a ReTM-Dictionary-based binaural signal denoising method that accounts for dynamic head movements. Our approach leverages a pre-computed dictionary of ReTMs corresponding to various head orientations. When the user first enters the room and takes a seat, natural head movements will cause the system to capture noise recordings from multiple orientations. These recordings are then used to compute a ReTM-Dictionary, mapping static noise characteristics to specific head poses. During the actual conversation, the AR glasses continuously track the user's head orientation in real time. Based on the detected orientation, the appropriate ReTM from the dictionary is used to effectively denoise the incoming noisy mixture.

3.1 Framework for ReTM Dictionary

This section builds upon our previous work in [25], where we introduced the use of ReTM for speech denoising in scenarios involving multiple simultaneously active sources and receivers. Here, we assumed that noise sources were continuously active, while speech sources appeared only intermittently. Additionally, both the noise sources and the recording microphones were considered spatially stationary, which allowed us to average the ReTM across time frames.

However, for applications such as AR glasses equipped with embedded microphone arrays, this assumption no longer holds. Head movements cause the orientation of the microphone array to change over time, resulting in spatial variations in the transfer functions between sources and microphone groups A and B.

To address the time variant requirement of the ReTM for applications having dynamic array orientations, we propose a ReTM dictionary that captures different array orientations corresponding to various head positions. Assume we have access to multi-channel (≥ 3) recordings for static noise sources but varying head orientations, we can segment the recordings into intervals during which the head orientation remains constant. For each such interval, we estimate a ReTM using (6), thereby building a dictionary of ReTMs associated with different orientations.

3.2 Binaural signal denoising

For binaural signal denoising using the ReTM dictionary, we choose the binaural channels as the target group $\{A\}$, and the remaining channels on the array as the reference group $\{B\}$. Consider a reverberant environment with $\mathcal{L} = \mathcal{L}_{\mathrm{S}} + \mathcal{L}_{\mathrm{N}}$ sound sources, where $\mathcal{L}_{\mathrm{S}}$ represents speech sources and $\mathcal{L}_{\mathrm{N}}$ represents noise sources. We also define source signal $\boldsymbol{S}(f,t) \triangleq [\boldsymbol{S}^{(\mathrm{S})}; \boldsymbol{S}^{(\mathrm{N})}]^T$, where $\boldsymbol{S}^{(\mathrm{S})}(f,t)$ represents speech signals of dimension $\mathcal{L}_{\mathrm{S}} \times 1$,

and $\boldsymbol{S}^{(\mathrm{N})}$ represents noise signals of dimension $\mathcal{L}_{\mathrm{N}} \times 1$. Also, $\boldsymbol{H}_{\mathrm{A}}(f,\theta,\phi) = [\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})}\ \boldsymbol{H}_{A}^{(\mathrm{N})}]$ and $\boldsymbol{H}_{\mathrm{B}}(f,\theta,\phi) = [\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}\ \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}]$ are the matrix with elements defined by the acoustic transfer functions. Here, (θ,ϕ) are the azimuth and elevation of the microphone array orientation. Also, $\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})}$, $\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}$ are the transfer functions from speech sources to microphone groups $\{A\}$ and $\{B\}$, and $\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}$, $\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}$ are the transfer functions from noise sources to microphone groups $\{A\}$ and $\{B\}$.

Hence, from (4) the time variant noise source ReTM of microphone groups $\{A\}$ and $\{B\}$ is,

$$\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi) \triangleq \boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}(f,\theta,\phi)(\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}(f,\theta,\phi))^{\dagger}, \quad (7)$$

where $(\cdot)^{\dagger}$ denotes Moore–Penrose inverse, assuming it is valid.

To perform binaural denoising, we begin by determining the head orientation for each recording segment using head-tracking data from AR glasses. Based on the identified orientation, we select the appropriate ReTM matrix ($\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi)$) from the precomputed dictionary. This matrix is then multiplied by the signal vector from microphone group B, denoted as $\boldsymbol{M}_{\mathrm{B}}(f,t)$, followed by substracting the result from $\boldsymbol{M}_{\mathrm{A}}(f,t)$ to obtain,

$$\begin{aligned}&\boldsymbol{M}_{\mathrm{A}}(f,t) - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi)\ \boldsymbol{M}_{\mathrm{B}}(f,t)\\ &= [\boldsymbol{H}_{\mathrm{A}}(f,\theta,\phi) - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi)\ \boldsymbol{H}_{\mathrm{B}}(f,\theta,\phi)]\ \boldsymbol{S}(f,t),\end{aligned} \quad (8)$$

where we use (1) and (2). Following the approach in [24], we substitute the expression for $\boldsymbol{H}_{\mathrm{A}}(f,\theta,\phi)$, $\boldsymbol{H}_{\mathrm{B}}(f,\theta,\phi)$, $\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi)$, and $\boldsymbol{S}(f,t)$ in (8), leading to the denoised signal:

$$\begin{aligned}&\boldsymbol{M}_{\mathrm{A}}(f,t) - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi)\ \boldsymbol{M}_{\mathrm{B}}(f,t)\\ &= [\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})}(f,\theta,\phi) - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f,\theta,\phi)\ \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}(f,\theta,\phi)]\ \boldsymbol{S}^{(\mathrm{S})}(f,t)\end{aligned} \quad (9)$$

which represents the spatially filtered speech component based on orientation-specific ReTM.

4. SIMULATION ANALYSIS

In this section, we evaluate the effectiveness of ReTM-based binaural denoising across different SNR levels and ReTM dictionary resolutions. The resolution of the ReTM dictionary is defined by the step size of the azimuth angle (θ) within the range $-45°$ to $45°$, where a larger step size corresponds to a lower resolution. The performance of the binaural denoising algorithm is assessed through numerical simulations using the SMIR generator [26]. We use five widely adopted evaluation metrics: Short-Time Objective Intelligibility (STOI) [28] and Perceptual Evaluation of Speech Quality (PESQ) [29] for speech clarity, Segmental SNR (SegSNR) [30, 31] for noise suppression capability, and Interaural Time Difference (ITD) and Interaural Level Difference (ILD) [32] for preserving binaural cues. STOI scores range from 0 to 1, with higher values indicating better speech intelligibility. PESQ scores range from -0.5 to 4, with higher scores reflecting better perceived speech quality. Similarly, higher SegSNR values indicate more effective noise suppression.

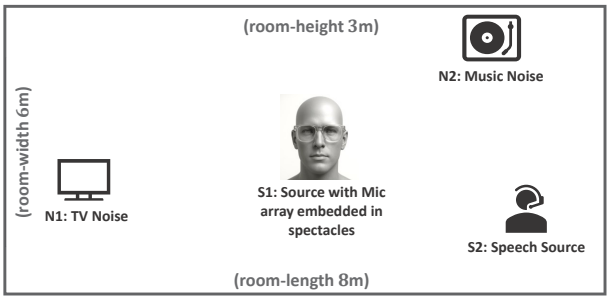

Figure 2: Approximate setup of the receiver and source positions in the SMIR based numerical simulations.

4.1 SMIR based Numerical Simulation

For the numerical simulations, we model the acoustic transfer function between sources and receivers using an ISM-based SMIR generator [26, 33]. The simulation takes place in a rectangular room with dimensions 8 m × 6 m × 3 m, and a reverberation time (RT60) of 645 ms. The SMIR generator is configured for a rigid sphere model with 30 harmonics. As shown in Fig.2, the microphone array is positioned on the head of S1 at coordinates $(4.0, 2.5, 1.0)$ m. The target speech source (S2) is positioned 1.45m from the center of the microphone array, in the same horizontal plane, with an azimuth angle of $40°$. Two interference sources (N1 and N2) are placed 1.85m and 1.20m from the microphone array center, also in the same horizontal plane, at azimuth angles $150°$ and $310°$ respectively. To simulate head rotations for S1, we vary the azimuth angle (θ) within the range $-45°$ to $45°$, while keeping the elevation fixed at $\phi = 0$. All signals are processed in the short-time Fourier domain using a window size of 16384 samples, a sampling rate of 16 kHz, and a

total duration of 300 seconds. We simulate two interfering source scenarios: i) 1 Noise (Music (N2)), 1 Speech (S2) - 1N1S, ii) 2 Noise (TV (N1), Music (N2)), 1 Speech (S2) - 2N1S, with SNR in the -10 dB: 5 dB: 10 dB range. In addition, to simulate thermal noise at the microphones, Additive White Gaussian Noise (AWGN) with a SNR of 40 dB is added to each source signal.

We generate two sets of recordings: training and test. For the training set, we simulate separate recordings for each source configurations (1N1S and 2N1S), and discrete head orientation (θ : $-45°$ to $45°$, $\phi = 0$), using pre-computed room impulse responses (RIRs). In these recordings, only the noise sources are active, while the speech source remains silent. Each head orientation is recorded for 60 seconds, with the total number of recordings depending on the ReTM resolutions. These recordings are used to compute the ReTM dictionary, as described in sub-section 2.3.

For the test set, we simulate recordings in which the head orientation dynamically changes among 8 to 12 randomly selected azimuth angles within the same $-45°$ to $45°$ range. For each SNR level and interference condition (1N1S and 2N1S), we generate 300 seconds of audio. During this period, the head orientation can change multiple times, with each orientation maintained for a minimum of 5 seconds. The noise sources remain continuously active throughout, while the speech source is intermittently active and consists of female utterances drawn from the TIMIT dataset [34]. Head orientation is tracked every 0.5 seconds and used to select the appropriate ReTM from the dictionary during denoising, as described in subsection 3.2.

4.2 Results and Discussions

Table 1: Comparison of Binaural Signal Cues across various ReTM Resolutions for 1N1S, SNR level 0 dB

ReTM Resolution (w.r.t θ step-size)	ITD (ms)		ILD (dB)	
	Before	After	Before	After
Ref. ($\theta = 0°$)	0.227	0.295	-0.458	-1.127
45°		0.283		-0.926
30°		0.261		-0.673
15°		0.244		-0.581
10°		0.238		-0.517
5°		0.233		-0.482
2.5°		0.231		-0.479

Table 1 compares the ITD and ILD scores for various ReTM-Dictionary resolutions, in a 1N1S - interfering source scenario at 0 dB SNR. The reference corresponds to the reverberant speech signal at θ = 0°, while the other rows show post-denoising values for increasing ReTM-Dictionary resolutions (i.e., smaller θ step-size). The results demonstrate that higher ReTM resolutions lead to improved preservation of binaural cues. As the step-size of θ decreases (from 45° to 2.5°), both ITD and ILD values after denoising increasingly align with the reference. For instance, at $\theta = 2.5°$ (ReTM resolution = 37), the deviation in ITD and ILD scores before and after denoising is minimal. In contrast, at $\theta = 45°$ (ReTM resolution = 3), the post-denoising ITD and ILD scores show significant deviation, indicating notable loss of spatial information. These findings are consistent with our expectations, as higher ReTM resolutions allow more accurate estimation of the noise field at the binaural channels (microphone group $\{A\}$), enabling better preservation of spatial cues during denoising.

Table 2 analyses the impact of SNR levels and ReTM-Dictionary resolutions (in terms of θ step-size) by comparing the speech quality metrics - STOI, PESQ, and SegSNR scores before and after denoising for interfering source scenario - 2N1S. From the table, it is evident that STOI, PESQ, and SegSNR scores improve with an increase in ReTM resolution (i.e., smaller θ step-size). This improvement is most notable at lower SNR levels (-10 dB), as the enhancement effects are more significant. For instance, at -10 dB, the STOI score improves from 0.357/0.335 (Left/Right) at the reference to 0.475/0.462 at 2.5° θ step-size, and PESQ score improves from 1.457/1.417 to 1.818/1.791. Similarly, segSNR score increases from -3.595/-3.890 to -2.746/-2.851, indicating the noise suppression effect. As the SNR level increases from -10 dB to +10 dB, the overall speech quality improves. However, the relative benefit of using a finer ReTM resolution decreases at higher SNRs, suggesting a saturation effect for low SNR levels. Additionally, the speech quality metrics—STOI, PESQ, and SegSNR—consistently yield high scores, demonstrating the effectiveness of the proposed binaural denoising algorithm. Overall, finer ReTM resolutions lead to better speech enhancement performance, and this makes sense as higher ReTM resolutions allow more accurate estimation of the noise field at the binaural channels (microphone group $\{A\}$), enabling better preservation of spatial cues.

Table 2: Comparison of Speech Quality Metrics across various ReTM Resolutions and SNR Levels for 2N1S

SNR	ReTM Resolution (θ step-size)	STOI Before		STOI After		PESQ Before		PESQ After		segSNR Before		segSNR After	
		Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
-10	Ref. (Az = 0°)	0.199	0.163	0.357	0.335	0.917	0.869	1.457	1.417	-9.631	-10.00	-3.595	-3.890
	45°			0.366	0.351			1.496	1.459			-3.457	-3.618
	30°			0.384	0.373			1.552	1.520			-3.296	-3.417
	15°			0.402	0.390			1.629	1.602			-3.119	3.296
	10°			0.440	0.417			1.715	1.697			-2.985	-3.058
	5°			0.467	0.434			1.786	1.758			-2.785	-2.875
	2.5°			0.475	0.462			1.818	1.791			-2.746	-2.851
0	Ref. (Az = 0°)	0.419	0.399	0.559	0.539	1.496	1.479	1.928	1.879	-0.093	-0.803	3.999	3.682
	45°			0.569	0.555			1.997	1.953			4.346	4.165
	30°			0.589	0.567			2.060	2.017			4.429	4.249
	15°			0.609	0.587			2.144	2.101			4.371	4.317
	10°			0.639	0.620			2.228	2.185			4.528	4.465
	5°			0.679	0.662			2.289	2.249			4.675	4.586
	2.5°			0.694	0.690			2.317	2.298			4.694	4.652
10	Ref. (Az = 0°)	0.582	0.571	0.650	0.647	1.762	1.744	2.404	2.390	10.138	10.085	12.209	11.965
	45°			0.675	0.657			2.457	2.427			12.261	12.042
	30°			0.698	0.676			2.520	2.479			12.413	12.298
	15°			0.726	0.705			2.569	2.538			12.566	12.397
	10°			0.751	0.741			2.597	2.572			12.706	12.579
	5°			0.795	0.779			2.637	2.630			12.928	12.794
	2.5°			0.803	0.791			2.642	2.640			12.986	12.879

5. CONCLUSION

In this paper, we have presented a binaural signal denoising algorithm leveraging ReTM dictionary as a spatial feature for head-mounted microphone array applications. By constructing a dictionary of noise-only ReTMs corresponding to various head orientations and leveraging head-tracking data in real time, our approach effectively adapts to dynamic head movements, crucial for AR or head-mounted microphone array applications. The results demonstrate that higher ReTM-Dictionary resolutions lead to improved preservation of binaural cues (ITD and ILD). The speech quality metrics (STOI, PESQ, SegSNR) also show consistently high scores, indicating the effectiveness of the proposed binaural denoising algorithm. Overall, the results are consistent with our expectations, as higher ReTM resolutions allow more accurate estimation of the noise field at the binaural channels (microphone group $\{A\}$), enabling better preservation of spatial cues during denoising. Future extensions involve developing a machine learning-based model that emulates our algorithm and evaluating its performance in dynamic noise environments with multiple simultaneous target sources.

6. REFERENCES

[1] P. Derleth, E. Georganti, M. Latzel, G. Courtois, M. Hofbauer, J. Raether, and V. Kuehnel, "Binaural signal processing in hearing aids," in *Seminars in Hearing*, vol. 42, pp. 206–223, Thieme Medical Publishers, Inc., 2021.

[2] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, 2007.

[3] N. Yousefian, P. C. Loizou, and J. H. Hansen, "A coherence-based noise reduction algorithm for binaural hearing aids," *Speech Communication*, vol. 58, pp. 101–110, 2014.

[4] F. Henry, M. Glavin, and E. Jones, "Noise reduction in cochlear implant signal processing: A review and recent developments," *IEEE reviews in biomedical engineering*, vol. 16, pp. 319–331, 2021.

[5] A. Zedan, T. Jürgens, B. Williges, B. Kollmeier, K. Wiebe, J. Galindo, and T. Wesarg, "Speech intelligibility and spatial release from masking improvements using spatial noise reduction algorithms in bimodal cochlear implant users," *Trends in Hearing*, vol. 25, p. 23312165211005931, 2021.

[6] R. Gupta, J. He, R. Ranjan, W.-S. Gan, F. Klein, C. Schneiderwind, A. Neidhardt, K. Brandenburg, and V. Välimäki, "Augmented/mixed reality audio for hearables: Sensing, control, and rendering," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 63–89, 2022.

[7] S.-N. Yao, "Headphone-based immersive audio for virtual reality headsets," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 300–308, 2017.

[8] W. Jin, M. J. Taghizadeh, K. Chen, and W. Xiao, "Multi-channel noise reduction for hands-free voice communication on mobile phones," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 506–510, IEEE, 2017.

[9] H.-W. Gierlich, "Binaural analysis methods and their relationship to quality evaluation of hands-free telecommunication equipment," in *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, pp. 1–4, IEEE, 1996.

[10] M. Tammen and S. Doclo, "Deep multi-frame mvdr filtering for binaural noise reduction," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, IEEE, 2022.

[11] M. Zohourian and R. Martin, "Gsc-based binaural speaker separation preserving spatial cues," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 516–520, IEEE, 2018.

[12] E. Hadad, S. Doclo, and S. Gannot, "The binaural lcmv beamformer and its performance analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, 2016.

[13] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function mvdr beamformers with interference cue preservation constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2449–2464, 2015.

[14] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[15] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2162–2176, 2015.

[16] B. Cornelis, M. Moonen, and J. Wouters, "Reduced-bandwidth multi-channel wiener filter based binaural noise reduction and localization cue preservation in binaural hearing aids," *Signal processing*, vol. 99, pp. 1–16, 2014.

[17] D. Marquardt, V. Hohmann, and S. Doclo, "Coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8648–8652, IEEE, 2013.

[18] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2384–2397, 2015.

[19] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 27, no. 2, pp. 457–468, 2018.

[20] A. Somayazulu, C. Chen, and K. Grauman, "Self-supervised visual acoustic matching," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24349–24367, 2023.

[21] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[22] T. May, "Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 406–414, 2017.

[23] B. Rafaely, V. Tourbabin, E. Habets, Z. Ben-Hur, H. Lee, H. Gamper, L. Arbel, L. Birnie, T. Abhayapala, and P. Samarasinghe, "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes–review and challenges," *Acta Acustica*, vol. 6, p. 47, 2022.

[24] T. Abhayapala, L. Birnie, M. Kumar, D. Grixti-Cheng, and P. Samarasinghe, "Generalizing the Relative Transfer Function to a Matrix for Multiple Sources and Multichannel Microphones," in *Eur. Signal Process. Conf. (EUSIPCO)*, IEEE, 2023.

[25] M. Kumar, L. Birnie, T. Abhayapala, A.-H. S., B. A., D. Grixti-Cheng, and P. Samarasinghe, "Speech denoising in multi-noise source environments using multiple microphone devices via Relative Transfer Matrix," in *Eur. Signal Process. Conf. (EUSIPCO)*, IEEE, 2024.

[26] D. P. Jarrett, E. A. Habets, M. R. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[27] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE trans. Antennas and Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Int. Conf. Acoust., Speech and Signal Process.(ICASSP)*, pp. 4214–4217, IEEE, 2010.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Int. Conf. Acoust., Speech, and Signal Process.(ICASSP)*, vol. 2, pp. 749–752, IEEE, 2001.

[30] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint arXiv:2107.04174*, 2021.

[31] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An introduction to the speech enhancement for augmented reality (spear) challenge," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, IEEE, 2022.

[32] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *the Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, 1977.

[33] E. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[34] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.