Speech denoising in multi-noise source environments using multiple microphone devices via Relative Transfer Matrix

Manish Kumar*, Lachlan Birnie, Thushara Abhayapala, Sandra Arcos Holzinger,
Amy Bastine, Daniel Grixti-Cheng, Prasanga Samarasinghe
Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, Australia
*E*mail*: Manish.Kumar@anu.edu.au

***Abstract*—Speech denoising is a challenging problem when there are multiple active noise sources. This paper introduces a novel blind denoising approach using the Relative Transfer Matrix (ReTM) as a spatial feature of noise source locations and the environment in multi-microphone settings. The ReTM is a generalization of Relative Transfer Function (ReTF) for simultaneously active sources and multiple receivers. We allocate receivers into two multichannel groups and formulate the ReTM to describe the spatial mapping between them. The ReTM with respect to noise sources is estimated blindly using covariance matrices of microphone recordings during speech-free intervals. We use the ReTM to estimate the noise at one group of microphones from the other. The estimated noise is then subtracted from the incoming signal to achieve speech denoising. We illustrate the effectiveness of the proposed algorithm through simulations and experimental recordings. The method does not require prior knowledge of the number of speech and noise sources, nor source and microphone locations, and can be extended to a configuration with more than three microphones.**

***Index Terms*—Relative Transfer Matrix, Denoising, Multi-Channel, STFT Domain, VAD**

I. Introduction

Speech enhancement and denoising algorithms are crucial in various applications, including video-conferencing, hands-free devices [1], hearing aid technologies [2], and robust automatic speaker recognition systems [3], [4]. Several denoising algorithms using single or multiple microphones have been developed over the past years, and they can be mainly categorized into non-parametric and parametric methods. Parametric algorithms relying on statistical signal generation models often suffer from assumptions on signal models, challenges in parameter estimation, inflexibility to diverse noise types, and potential computational overhead [5]–[8]. On the other hand, non-parametric techniques like spectral subtraction [9], non-negative matrix factorization [10], and wavelet denoising [11] are commonly used for speech enhancement. However, these methods do not consider blind denoising in environments with multiple active noise sources. Recent works using deep learning methods [12], [13] have explored such scenarios, but these approaches require large training data and exhibit performance degradation when faced with noise types differing from their trained characteristics.

In this study, we introduce a non-parametric denoising approach that leverages the Relative Transfer Matrix (ReTM) [14] of multiple noise sources. We build upon the theoretical framework of ReTM [14], specifically applying it to the practical domain of speech denoising. The ReTM describes the coupling of acoustic channels between two groups of multi-channel receivers [14], preserving three key properties of the ReTF: $i)$ independence from the source signals, $ii)$ distinctive spatial characteristics determined by the source-microphone placements and the surrounding environment, and $iii)$ the ability for blind estimation from observed signals [14].

We anticipate that the ReTM's versatility will broaden the scope of its acoustic applications, previously explored through ReTF, to include complex multi-source environments. These applications encompass source separation [15], [16], beamforming [17], and speech denoising [4]. Notably, ReTF was shown to achieve strong speech denoising for a single interfering noise source [4].

Exploiting the ReTM as a spatial feature requires the assumption that noise sources and recording microphones remain spatially stationary. This assumption holds true in a variety of real-world situations. For example, in buildings, noises from heating and cooling machinery remain relatively stationary to mounted microphones. Likewise, vehicle hands-free devices are stationary compared to engine and air-conditioning noise.

In this paper, we propose a speech-denoising algorithm that utilizes blind estimation to derive the ReTM of multiple noise sources from covariance matrices of recorded signals, potentially distributed across multiple devices. Here, we assume that a segment of the recording containing only the noise sources (no active speech) is available and can be isolated through Voice Activity Detection (VAD). Subsequently, we utilize the estimated ReTM to determine the noise in a group of microphones we denote as the reference microphone group. The estimated noise is then subtracted from the received signal, effectively achieving speech denoising. The presented results validate the effectiveness of our approach through both simulated and live recordings under low signal-to-noise ratio (SNR) conditions involving more than three noise sources.

The subsequent sections of this paper are structured as follows: Section 2 presents the formulation of the complex

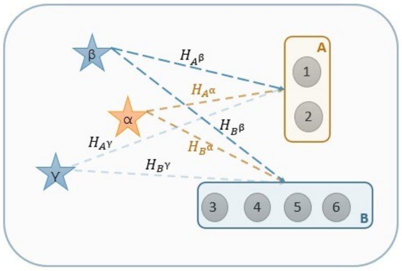

Fig. 1. Illustration of grouped microphones. Here, $\{A\}$ and $\{B\}$ denotes microphones subgroups and $\{\alpha, \beta, \gamma\}$ denotes sources.

noisy mixture that we aim to denoise. Section 3 elaborates on our denoising methodology, which involves the proposed use of the ReTM to spatially model noise sources. In Section 4, we examine the applicability of ReTM in speech denoising for both simulated and live recordings. We also compare the performance of the proposed model with the established benchmark of the Multi-Channel Wiener Filter (MWF) [18]. Lastly, in Section 5, we conclude by summarizing the key findings and outlining directions for future research.

II. System Model

Consider a reverberant environment with Q microphones (potentially across multiple devices in a sensor network) with the received signals in the Short Time Fourier Transform (STFT) domain $M_q(f,t)$, $q = \{1,\cdots,Q\}$ and $\mathcal{L}$ sound (including both speech and noise) sources $S_\ell(f,t)$, $\ell = \{1,\cdots,\mathcal{L}\}$. Let us separate the microphones into two subgroups (reference and target) denoted by $\{A\}$ and $\{B\}$, assigned with Q_A and Q_B microphones, respectively. We express the signals received by each microphone group in matrix form as

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

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

where $\boldsymbol{M}_{\mathrm{A}} = [M_1,\cdots,M_{Q_A}]^T$, $\boldsymbol{S} = [S_1,\cdots,S_{\mathcal{L}}]^T$, $[\cdot]^T$ 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. The vector $\boldsymbol{M}_{\mathrm{B}} \in \mathbb{C}^{Q_B \times 1}$ and the matrix $\boldsymbol{H}_{\mathrm{B}} \in \mathbb{C}^{Q_B \times \mathcal{L}}$ are similar. Note that we omit thermal microphone noise in the above formulation for an easier explanation of the theoretical development in the next section.

III. Speech Denoising with Noise ReTM

In this section, we briefly introduce ReTM as a spatial feature that is a generalization of ReTF to multiple simultaneous sound sources, in this case noise sources. We then mathematically show how a multiple microphone recordings of noisy speech can be denoised provided the noise source ReTM is known.

Here, we consider noise sources to be continuously active, while speech sources are present only sporadically. Our objective is to denoise the recording enough to make the content of the processed signal understandable to listeners. While we look at denoising speech in this paper, the proposed method could be applied to spatially denoise any non-continuous signal.

A. Relative Transfer Matrix

The ReTM describes the spatial mapping between two microphone groups in response to the given sound sources [14]. As an illustration, let us assume a reverberant environment (see Figure 1) with $Q = 6$, $Q_A = 2$, $Q_B = 4$, and $\mathcal{L} = 3$ where $\ell = \{\alpha, \beta, \gamma\}$. Also, let α represent a single speech source (S_p), and $\{\beta, \gamma\}$ represents two unmoving noise sources ($\boldsymbol{S}^{(\mathrm{N})}$). Hence, the ReTM, $\boldsymbol{\mathcal{R}}_{AB}(f)$, between the microphone groups - $\{A\}$ and $\{B\}$ in response to given sound sources $\ell = \{\alpha, \beta, \gamma\}$ can be modeled as,

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

The theoretical definition of the ReTM [14] 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

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

where $(\cdot)^\dagger$ denotes Moore–Penrose inverse, assuming it is valid. Here, we observe that the source signals elegantly cancel off, leaving behind a spatial feature of the sources. Hence, the ReTM (4) is a matrix defined solely by the spatial properties (the transfer functions) of the sound sources [14].

B. Blind estimation of Noise Source ReTM

In order to estimate the Noise Source ReTM, we need a segment of recordings that contain only the noise signals denoted by $\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})}$ and $\boldsymbol{M}_{\mathrm{B}}^{(\mathrm{N})}$. These can be extracted using an appropriate VAD such as [19], or by using ground truth data. In this study, we utilized ground truth data to identify and extract the noise-only segments of the recordings. In existing work, there are several methods to directly estimate the ReTF using noise-only segments [20]. Here, we exploit a covariance matrices-based approach [14], where covariance matrices of microphones groups $\{A\}$ and $\{B\}$ can be defined as,

$$\boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f) \triangleq E\Big\{\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})}\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})^*}\Big\} \text{ and } \boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f) \triangleq E\Big\{\boldsymbol{M}_{\mathrm{B}}^{(\mathrm{N})}\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})^*}\Big\} \tag{5}$$

where, $[\cdot]^*$ is conjugate transpose, and $E\{\cdot\}$ denotes the expectation which can be estimated from averaged time frames,

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f) &\cong \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})}(f,t)\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})^*}(f,t) \\ \boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f) &\cong \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\boldsymbol{M}_{\mathrm{B}}^{(\mathrm{N})}(f,t)\boldsymbol{M}_{\mathrm{A}}^{(\mathrm{N})^*}(f,t). \end{aligned} \tag{6}$$

Using (1) and (2) in (5), we write

$$\boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f) = \boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}\boldsymbol{\mathcal{P}}_{\boldsymbol{S}}^{(\mathrm{N})}\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})^*} \tag{7}$$

$$\boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f) = \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}\boldsymbol{\mathcal{P}}_{\boldsymbol{S}}^{(\mathrm{N})}\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})^*} \tag{8}$$

where $\boldsymbol{\mathcal{P}}_{\boldsymbol{S}}^{(\mathrm{N})} \triangleq E\{\boldsymbol{S}^{(\mathrm{N})}\boldsymbol{S}^{(\mathrm{N})^*}\}$ and $\boldsymbol{S}^{(\mathrm{N})}$ is the vector of noise signals. By multiplying (8) by $\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})^\dagger}$ and substituting into (7), we obtain

$$\boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f) = \boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})^\dagger}\boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f) = \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f)\,\boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f) \tag{9}$$

The ReTM is estimated by applying the pseudo-inverse of $\boldsymbol{\mathcal{P}}_{\mathrm{BA}}(f)$ to (9) as

$$\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}(f) = \boldsymbol{\mathcal{P}}_{\mathrm{AA}}(f)\boldsymbol{\mathcal{P}}_{\mathrm{BA}}^{\dagger}(f). \tag{10}$$

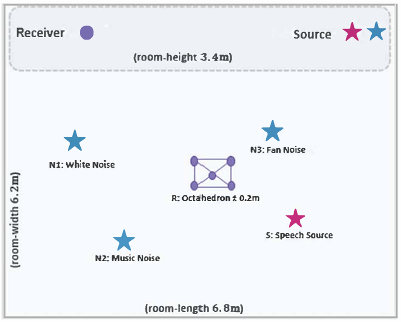

Fig. 2. Approximate setup of the receiver and source positions in the numerical simulation and experimental recordings.

In practice, microphones will have additive thermal noise, thus the estimation of ReTM in (10) is only an approximation. However, the main advantage of ReTM based approach is that it does not require any prior knowledge of the number of sound sources, unlike in [21]. Furthermore, the estimation of ReTM relies on covariance matrices, a methodology akin to the widely used MUSIC source localization technique that has been used in various acoustic applications [14] [22].

C. Speech denoising using ReTM of noise sources

Let there be $\mathcal{L}_{\mathrm{S}}$ speech and $\mathcal{L}_{\mathrm{N}}$ noise sources. We define $\boldsymbol{S}(f,t) \triangleq [\boldsymbol{S}^{(\mathrm{S})}(f,t); \boldsymbol{S}^{(\mathrm{N})}(f,t)]^T$, where $\boldsymbol{S}^{(\mathrm{S})}(f,t)$ represents speech signals of dimension $\mathcal{L}_{\mathrm{S}} \times 1$, and $\boldsymbol{S}^{(\mathrm{N})}(f,t)$ represents noise signals of dimension $\mathcal{L}_{\mathrm{N}} \times 1$. Additionally, we define $\boldsymbol{H}_{\mathrm{A}} = [\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})} \;\; \boldsymbol{H}_{A}^{(\mathrm{N})}]$ and $\boldsymbol{H}_{\mathrm{B}} = [\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})} \;\; \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}]$. Here, $\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})}$ and $\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}$ are the transfer functions from speech sources to microphone groups $\{A\}$ and $\{B\}$, and $\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}$ and $\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}$ are the transfer functions from noise sources to microphone groups $\{A\}$ and $\{B\}$. Note that we drop f and t dependency from quantities for brevity.

Assume that we have estimated the ReTM of noise sources $\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}$ (as outlined in the preceding section) using a speech-free segment of recorded signals. To denoise the recorded signal containing speech, we multiply the signal vector from group $\{B\}$ denoted as $\boldsymbol{M}_{\mathrm{B}}$ by the estimated ReTM ($\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}$) and then subtract the result from $\boldsymbol{M}_{\mathrm{A}}$ to obtain

$$\boldsymbol{M}_{\mathrm{A}} - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, \boldsymbol{M}_{\mathrm{B}} = [\boldsymbol{H}_{\mathrm{A}} - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, \boldsymbol{H}_{\mathrm{B}}]\, \boldsymbol{S}, \tag{11}$$

where we use (1) and (2). By substituting for $\boldsymbol{H}_{\mathrm{A}}$, $\boldsymbol{H}_{\mathrm{B}}$, $\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}$, and $\boldsymbol{S}$ in (11), we have

$$\begin{aligned}
&\boldsymbol{M}_{\mathrm{A}} - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, \boldsymbol{M}_{\mathrm{B}} \\
&= [[\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})}\; \boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}] - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, [\boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}\; \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}]]\; [\boldsymbol{S}^{(\mathrm{S})}\, ; \boldsymbol{S}^{(\mathrm{N})}]^{\mathrm{T}} \\
&= [\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})} - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}]\, \boldsymbol{S}^{(\mathrm{S})} \; - \; [\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})} - \underbrace{\boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{N})}}_{\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{N})}}]\, \boldsymbol{S}^{(\mathrm{N})} \\
&= [\boldsymbol{H}_{\mathrm{A}}^{(\mathrm{S})} - \boldsymbol{\mathcal{R}}_{\mathrm{AB}}^{(\mathrm{N})}\, \boldsymbol{H}_{\mathrm{B}}^{(\mathrm{S})}]\, \boldsymbol{S}^{(\mathrm{S})}
\end{aligned} \tag{12}$$

From (12), it becomes evident that the noise signals have been eliminated through the ReTM filtering and subtraction process. Additionally, any distortion left in the denoised speech signal is solely attributed to the room transfer matrix of both the speech and noise sources.

IV. Experimental Validation

In this section, we explore the applicability of ReTM in multi-channel speech denoising for various scenarios of three to five simultaneous noise sources and six microphones as shown in Fig. 2. We also evaluate the performance of the speech denoising algorithm for both Image Source Model (ISM) based simulated recording and experimental recording from live loudspeakers. Our analysis employed several widely used performance metrics for speech denoising, namely STOI [23], ESTOI [24], PESQ [25], SNR, and SDR [26]. For STOI, and ESTOI, a higher score indicates better speech intelligibility, while a higher PESQ score indicates better speech quality. Likewise, for SNR and SDR, a higher score represents superior audio fidelity.

A. Numerical Simulation

For experimental simulation, we modeled the acoustic transfer functions between the sources and receivers using the ISM [27]. A 7th order image depth and 0.8544 reflection coefficient on all walls were used ($T_{60} = 500$ ms). The six receivers were configured as an octahedron ±0.2 m along each axis. The two receivers on the z-axis comprised Group A, and the remaining were Group B. The signals were processed directly in the short-time Fourier domain for an 8192 window size, 16 kHz sampling rate, and 60-second duration. In our simulations, we considered three distinct scenarios: $i)$ 2 Noise sources (White, Music), 1 Speech - 2N1S, $ii)$ 3 Noise sources (White, Music, Fan), 1 Speech - 3N1S, and $iii)$ 4 Noise sources (White, Music, Fan, Motor), 1 Speech - 4N1S. Across all these scenarios, the noise sources were continuously active for the entire 60-second duration, whereas the speech source was active intermittently. For these simulations, we utilized female speech utterances sourced from the TIMIT dataset [28].

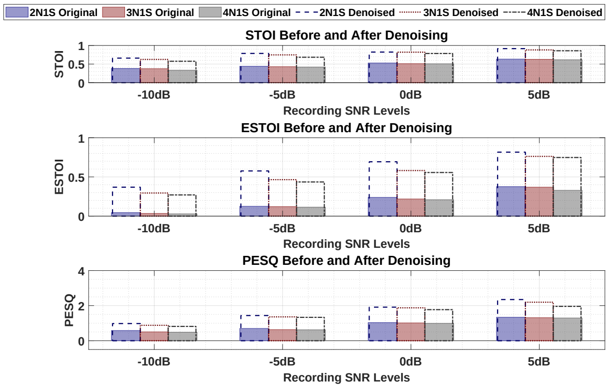

Fig. 3. STOI, ESTOI, and PESQ plot for denoised speech using the proposed method with ISM-based simulated recordings.

Fig. 3 compares the ISM simulated recordings with their denoised versions using STOI, ESTOI, and PESQ. The denoised recordings exhibit STOI scores from 0.62 to 0.91, ESTOI scores from 0.27 to 0.82, and PESQ scores from 0.81 to 2.34, indicating better speech quality and intelligibility. Furthermore, across all three scenarios (2N1S, 3N1S, and 4N1S), the denoised recordings consistently outperform their original

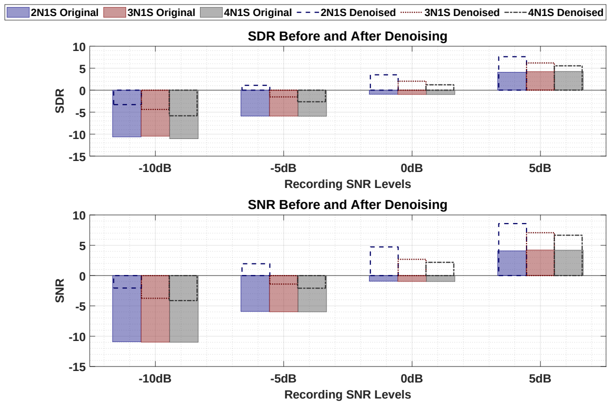

Fig. 4. SDR and SNR plot for denoised speech using the proposed method with ISM-based simulated recordings.

counterparts (STOI: $\geq 39\%$ improvement, ESTOI: $\geq 208\%$ improvement, and PESQ: $\geq 52\%$ improvement), highlighting the effectiveness of our denoising algorithm.

Fig. 4 compares the SNR and SDR for ISM simulated recordings and their denoised counterparts. Here, the denoised recordings exhibit consistent improvement in SDR (2.15dB to 7.34dB) and SNR (2.46dB to 8.57dB) compared to their original counterparts, underscoring the algorithm's effective denoising capability. Notably, the algorithm's effectiveness remains unaltered even with the introduction of additional noise sources.

B. Experimental Recordings

The live recording is performed inside a carpeted room with a dropped ceiling and some partial acoustic treatment (T60 = 181 ms). The recording also had noteworthy background air conditioning noise (45 dBA noise floor). We used 6 condenser microphones configured as an octahedron ±0.2 m along each axis for recording with a sampling rate of 48 kHz. The sound sources were 3 or 4 live loudspeakers playing different noise signals and a speech signal, placed $\approx$ 1.5 m around the microphone array, as shown in Fig.2. Similar to numerical simulation, we considered the first two scenarios of 2N1S and 3N1S.

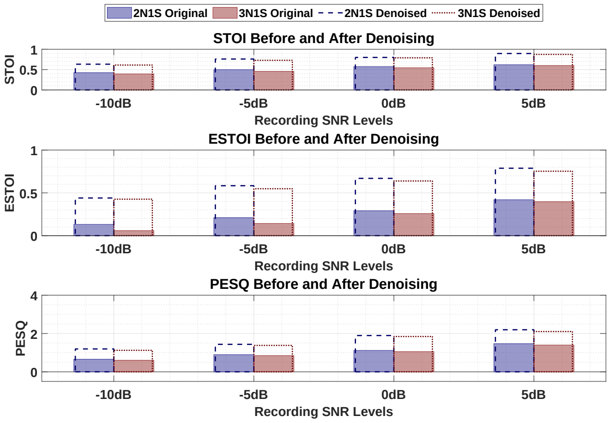

Fig. 5. STOI, ESTOI, and PESQ plot for denoised speech using the proposed method with live recordings.

In Fig. 5, the STOI, ESTOI, and PESQ scores of live recordings are compared with their denoised versions. Here also, the denoised recordings exhibit STOI scores from 0.61 to 0.89, ESTOI scores from 0.42 to 0.78 , and PESQ scores from 1.12 to 2.19. Additionally, the denoised recordings consistently outperform their live-recording counterparts (STOI: $\geq$ 44% improvement, ESTOI: $\geq$ 192% improvement, and PESQ: $\geq$ 54% improvement), indicating good speech quality and intelligibility.

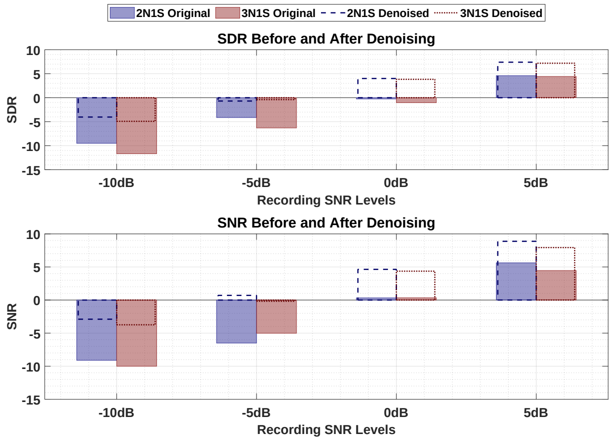

Fig. 6. SDR and SNR plot for denoised speech using the proposed method with live recordings.

Additionally, in Fig.6, we compare the SNR and SDR scores for live recordings with their denoised counterparts. Here also, the denoised recordings exhibit notable improvement in SDR (2.81dB to 6.94dB) and SNR (3.51dB to 6.22dB) from their original counterparts, highlighting the algorithm's effective denoising capability, even for live recordings.

C. Comparison with the baseline method

We now compare the performance of the ReTM-based denoising method with the MWF proposed in [18]. The prime reasons for choosing MWF [18] as a baseline for comparison are its multi-channel approach, and the filter weights being estimated blindly from the observed signals using covariance matrices. For brevity, we have only presented the comparison results for ISM simulated recordings with the 2N1S scenario.

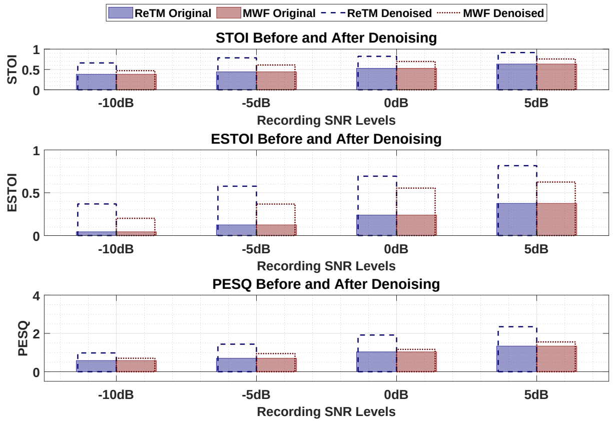

Fig. 7. STOI, ESTOI, and PESQ plot for denoised speech using the proposed and baseline [18] method with ISM-based simulated recordings.

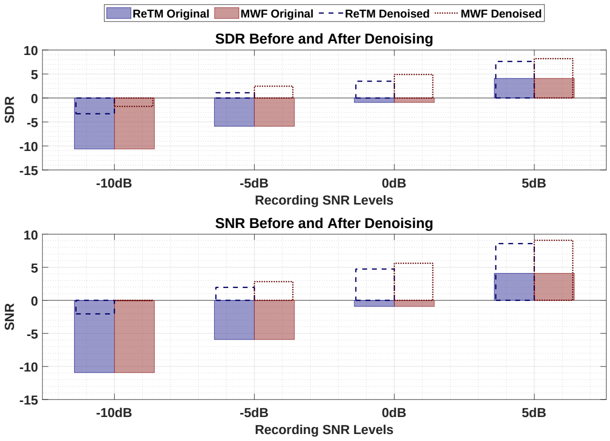

Fig. 8. SDR and SNR plot for denoised speech using the proposed the proposed and baseline [18] method with ISM-based simulated recordings

In Fig.7, the STOI, ESTOI, and PESQ scores of the ReTM-based denoising algorithm are compared against the baseline MWF [18] method. The proposed algorithm (ReTM) outperforms the baseline (MWF [18]) by $\geq 6\%$ for STOI, $\geq 24\%$ for ESTOI, and $\geq 37\%$ for PESQ indicating better speech quality and intelligibility of the denoised signal.

Moreover, from Fig.8 it is advent that the SDR and SNR scores of the proposed algorithm (ReTM) are comparable to those of the baseline (MWF [18]) method, demonstrating its effective denoising capability.

V. Conclusion

In this paper, we have introduced a robust multi-channel speech denoising algorithm utilizing the ReTM of the noise sources. The ReTM is estimated blindly using covariance matrices of microphone recordings during speech-free periods. The method does not require prior knowledge of the number of speech and noise sources, nor microphone locations, and can be extended to a configuration with more than three microphones. We have evaluated the performance of the proposed algorithm across various noise scenarios, including both simulated and live recordings. The results show consistently high STOI, ESTOI, and PESQ scores, along with satisfactory SDR and SNR, confirming its effective denoising capability. Future extensions involve developing a machine learning-based model that emulates our algorithm and exploring its performance in scenarios with a higher number of interfering noises than the reference microphone group.

References

[1] S. J. Park, C. G. Cho, C. Lee, and D. H. Youn, "Integrated echo and noise canceler for hands-free applications," *Trans. Circuits and Systems II: Analog and Digital Signal Process.*, vol. 49, no. 3, pp. 188–195, 2002.

[2] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Incorporating relative transfer function preservation into the binaural multi-channel wiener filter for hearing aids," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2016, pp. 6500–6504.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] L. Birnie, P. Samarasinghe, T. Abhayapala, and D. Grixti-Cheng, "Noise RETF Estimation and Removal for Low SNR Speech Enhancement," in *Proc. IEEE Workshop Mach. Learning Signal Process. (MLSP)*, 2021, pp. 1–6.

[5] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceed. of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[6] K. Chen, "On the use of different speech representations for speaker modeling," *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 3, pp. 301–314, 2005.

[7] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. on Inform. Theo.*, vol. 41, no. 3, pp. 613–627, 1995.

[8] S. Mallat et al., "A wavelet tour of signal processing: the sparce way," *AP Profess., Third Edi., London*, 2009.

[9] V. Vaseghi Saeed, "Advanced digital signal processing and noise reduction," *Wiley*, vol. 29, pp. 43, 2000.

[10] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP Journal on Adv. in Signal Process.*, vol. 2014, pp. 1–24, 2014.

[11] Y. Shao and C.H. Chang, "A generalized time–frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 4, pp. 877–889, 2007.

[12] Anurag Kumar and Dinei Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," *arXiv preprint arXiv:1605.02427*, 2016.

[13] Nasir Saleem and Muhammad Irfan Khattak, "Deep neural networks for speech enhancement in complex-noisy environments," 2020.

[14] 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.

[15] N. Ito, S. Araki, and T. Nakatani, "Permutation-free clustering of relative transfer function features for blind source separation," in *Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2015, pp. 409–413.

[16] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2010.

[17] N. Gößling and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2019, pp. 416–420.

[18] Wageesha N Manamperi, Thushara D Abhayapala, Prasanga N Samarasinghe, and Jihui Aimee Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones using multichannel wiener filtering and gaussian-mixture based post-filtering," *Applied Acoustics*, vol. 216, pp. 109818, 2024.

[19] M. Van Segbroeck, A. Tsiartas, and S. S. Narayanan, "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice," in *INTERSPEECH*, 2013, pp. 704–708.

[20] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2015, pp. 544–548.

[21] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2015, pp. 419–423.

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

[23] 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)*. IEEE, 2010, pp. 4214–4217.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[25] 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)*. IEEE, 2001, vol. 2, pp. 749–752.

[26] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide–Revision 2.0," 2005.

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

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