Successive Speaker Relative Transfer Function Estimation Through Relative Transfer Matrix in Noisy Reverberant Environments

Wageesha N. Manamperi, Thushara D. Abhayapala
The Australian National University, Canberra, Australia
E-mail: {*wageesha.manamperi, thushara.abhayapala*}*@anu.edu.au*

Abstract—**Estimating the Relative Transfer Functions (ReTFs) of multiple speakers in a noisy reverberant environment are beneficial to signal processing applications of acoustic scene analysis, denoising, dereverberation, signal enhancement, and separation. Whilst most audio applications exploit the covariance matrix structure of multichannel recordings, estimating ReTF in the presence of simultaneously active multiple speakers is still considered to be a challenging problem. In this paper, we propose a novel method for estimating the ReTF using the relative transfer matrix (ReTM) of multiple microphones for multi-talker scenarios, which is suitable for noisy reverberant rooms. The method is based on the noise-only ReTM, and calculates the covariance matrices of (i) the first speaker and noise, and (ii) both speakers and noise at two microphone groups to reconstruct the ReTF of the second speaker. We demonstrate the ReTF estimation accuracy using numerical simulation of two speakers and two noise sources in a reverberant environment. The proposed method offers an accurate estimation with a low Hermitian angle. Additionally, the proposed algorithm is shown to better extract the voice of the successive speaker from the noisy microphone recordings over various SNR levels using a minimum variance distortionless response beamformer with improved noise reduction performance.**

I. Introduction

A relative transfer function (ReTF) is a spatial feature that describes the acoustic channel between two microphones in response to a single sound source [1]. Knowledge of desired sound source ReTF is essential in many spatial signal processing applications, such as beamforming, source localization, speech enhancement, source separation, and acoustic echo cancellation [1]–[5]. As a result, ReTF estimation of the target speaker in one or more interfering speakers within a noisy reverbarant room is an active problem in audio signal processing.

Many approaches have been developed for estimating the ReTF of a single active sound source, such as least-squares [1], [6], [7], covariance subtraction [8], [9] and covariance whitening [10]–[12], as well as manifold learning [13]. Common to these approaches is the use of signals' power spectral density (PSD) matrices calculated during the (i) noise-only, and (ii) speech plus noise, time segments for estimating the ReTF. Despite this, multiple and concurrent speakers scenarios are challenging problems as it could instead provide joint subspace spanning of the ReTF of all speakers [10].

There exist some techniques that generalize the ReTF for multiple simultaneous sound sources [14], [15]. The recently proposed Relative Transfer Matrix (ReTM) [15], where receivers allocate into two multi-microphone groups, is seen to be well performed when applied to speech enhancement [16], [17], and speaker separation [18] in a multi-source noisy reverberant environment.

Several ReTF estimation methods in a multiple and concurrent speaker scenario have been proposed, including (i) an expectation maximization algorithm that assumes only single source active at a given time-frequency bin and noise PSD is time-invariant [19]; (ii) a joint diagonalization method with simultaneous confirmatory factor analysis [20]; (iii) a more robust joint diagonalization based on linear algebraic concepts [21]; (iv) an orthogonal Procrustes problem with a priori information of the ReTFs [22]; and (v) blind oblique projection method [23]. Common drawbacks to these approaches are assuming W-disjoint orthogonality conditions, and a sufficiently high signal-to-noise ratio (SNR).

In [24], Gode and Doclo introduce a covariance blocking and whitening method to estimate the ReTF of the successive speaker in a noisy reverberant environment from two speakers activate successively. Alternatively, this paper proposes a novel method for estimating the ReTF of the second speaker by utilizing some of the properties of the covariance matrices imposed by the ReTM within noisy reverberant rooms, thus neither the estimate of the noise covariance matrix nor the ReTF of the first speaker requires unlike in [24]. The derivation of this method is based on the sole assumption that first speaker and multiple noise sources do simultaneously active in a given time segment prior to the time segment of the dual-speaker plus all noise sources as in [24]. Similar to the ReTM, we derive the proposed method by dividing multiple microphones into two groups and calculating the covariance matrices between them for both the single-speaker and dual-speaker segments. In the sequel, we first formulate the problem and introduce the ReTM. Secondly, in Section III, we present the new ReTF estimation method. Then in Section IV, we verify this ReTF estimation algorithm of the second speaker via a simulation study and show increased similarity with the ground truth ReTF

Thanks to Australian Research Council (ARC) Discovery Project Grant DP200100693 for funding.

vector in both free-field and high reverberant environments at low SNR levels. Finally, in Section V, we further evaluate the noise reduction performance in terms of SNR improvement when employing the proposed method in a minimum variance distortionless response (MVDR) beamformer.

II. Problem Definition

In this section, we formulate the problem of estimating the Relative Transfer Function (ReTF) of the second speaker in a noisy mixture of two speakers that activate successively. We first present the system model for the specific scenario, and then review the Relative Transfer Matrix (ReTM), used by the proposed ReTF estimation algorithm.

A. System Model

Consider a reverberant environment with two speakers that are successively activated and $\mathcal{L}$ background noise sources. Let there be Q arbitrary distributed microphones. In the short time Fourier transform (STFT) domain, we express the received signal as $y_q(f,t), q=1,\cdots,Q$, and speech signal at each speaker as S_1 and S_2, and noise source signals as $N_\ell(f,t), \ell=1,\cdots,\mathcal{L}$. The received signals in matrix form as

$$\mathbf{y}(f,t) = \mathbf{h}_1(f)S_1(f,t) + \mathbf{h}_2(f)S_2(f,t) + \mathbf{H}_N(f)\mathbf{N}(f,t), \tag{1}$$

where $\mathbf{y}(f,t) = [y_1(f,t),\dots,y_Q(f,t)]^T$ and $\{\cdot\}^T$ is the matrix transpose, and $\mathbf{h}_1(f)$, and $\mathbf{h}_2(f)$ are the $[Q \times 1]$ vectors of the relative transfer functions (ReTFs) with respect to a reference microphone of the first and second speaker, respectively, $\mathbf{N}(f,t) = [N_1(f,t),\dots,N_{\mathcal{L}}(f,t)]^T$ and $\mathbf{H}_N(f)$ is the Q by $\mathcal{L}$ noise sources transfer function matrix, which we define next.

The problem discussed in this paper is to estimate the ReTF vector of the successive speaker, $\mathbf{h}_2(f)$ from the mixture of multichannel recordings, $\mathbf{y}(f,t)$ assuming a specific scenario of three time segments with known segment boundaries as (i) multiple noise sources-only segment ($\mathcal{T}_1$), (ii) first speaker and multiple noise sources ($\mathcal{T}_2$), and (iii) first and second speakers and multiple noise sources ($\mathcal{T}_3$) as shown in Fig. 1

B. The Relative Transfer Matrix

Consider two groups of microphones $\{A\}$ and $\{B\}$ assigned with Q_A and Q_B microphones, respectively ($Q = Q_A + Q_B$). $\mathbf{y}_A(f,t)$, and $\mathbf{y}_B(f,t)$ are denoted as the vector of microphone signals for each group, therefore, the received signals at each microphone group are given as

$$\mathbf{y}_A(f,t) = \mathbf{h}_{A1}(f)S_1(f,t)+\mathbf{h}_{A2}(f)S_2(f,t)+\mathbf{H}_{AN}(f)\mathbf{N}(f,t), \tag{2}$$

$$\mathbf{y}_B(f,t) = \mathbf{h}_{B1}(f)S_1(f,t)+\mathbf{h}_{B2}(f)S_2(f,t)+\mathbf{H}_{BN}(f)\mathbf{N}(f,t), \tag{3}$$

where $\mathbf{y}_A(f,t) = [y_A^{(1)}(f,t),\cdots,y_A^{(Q_A)}(f,t)]^T$, $\mathbf{h}_{A1}(f) = [h_{A1}^{(1)}(f),\dots,h_{A1}^{(Q_A)}(f)]^T$, $\mathbf{h}_{A2}(f) = [h_{A2}^{(1)}(f),\dots,h_{A2}^{(Q_A)}(f)]^T$, and $\mathbf{H}_{AN}(f)$ is a $[Q_A \times \mathcal{L}]$ matrix with elements defined by the acoustic transfer functions. The vectors $\mathbf{y}_B(f,t)$, $\mathbf{h}_{B1}(f)$, $\mathbf{h}_{B2}(f)$, and $\mathbf{H}_{BN}(f)$ are similarly defined.

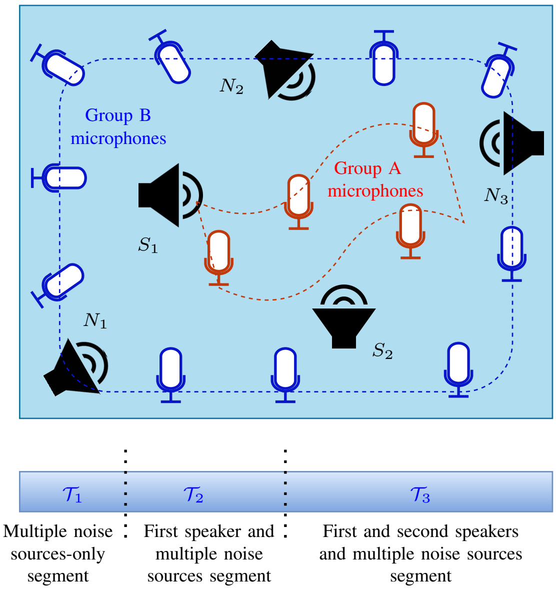
Group B
microphones
N_2
Group A
microphones
N_3
S_1
N_1
S_2
$\mathcal{T}_1$
$\mathcal{T}_2$
$\mathcal{T}_3$
Multiple noise sources-only segment
First speaker and multiple noise sources segment
First and second speakers and multiple noise sources segment

Fig. 1. Illustration of the problem setup.

The relative transfer matrix (ReTM) of noise sources, $\boldsymbol{\mathcal{R}}_{AB}^{(N)}(f)$, is defined as in [15]

$$\boldsymbol{\mathcal{R}}_{AB}^{(N)}(f) \triangleq \mathbf{H}_{AN}(f)\mathbf{H}_{BN}(f)^{\dagger}, \tag{4}$$

where $(\cdot)^\dagger$ is Moore-Penrose inverse, assuming the validity, i.e., $Q_B \geq \mathcal{L}$. Thus, at the first time segment $\mathcal{T}_1$, we can relate the received signal at group $\{A\}$ and $\{B\}$ using

$$\mathbf{y}_A^{(\mathcal{T}_1)}(f,t) = \boldsymbol{\mathcal{R}}_{AB}^{(N)}(f)\mathbf{y}_B^{(\mathcal{T}_1)}(f,t).$$

Note that ReTM is defined by the spatial properties of the sound sources such that it is independent of the sound source signals. In applications, the ReTM is constant for a stationary environment.

Assuming a stationary acoustic scenario, the next section shows how to estimate the ReTF of the successive speaker from the given multichannel recordings.

III. ReTF Estimation with ReTM

In this section, we propose a method for successive speaker ReTF estimation that utilizes some of the properties of the covariance matrices imposed by the ReTM.

A. Noise Sources ReTM

Here, we define the noise sources-only ReTM that will help for deriving our proposed method. Considering the background noise-only signals at $\mathcal{T}_1$ time segment, we can write the received signals as

$$\mathbf{y}_A^{(\mathcal{T}_1)} = \mathbf{H}_{AN}\mathbf{N}, \tag{5}$$

$$\mathbf{y}_B^{(\mathcal{T}_1)} = \mathbf{H}_{BN}\mathbf{N}. \tag{6}$$

The ReTM of the background noise sources can be estimated using the covariance matrices-based approach [15]. The background noise covariance matrices of microphone groups $\{A\}$ and $\{B\}$ are given as

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)}(f) &\triangleq E\{\mathbf{y}_A^{(\mathcal{T}_1)}(f,t)\mathbf{y}_A^{(\mathcal{T}_1)^*}(f,t)\}, \\ \boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_1)}(f) &\triangleq E\{\mathbf{y}_B^{(\mathcal{T}_1)}(f,t)\mathbf{y}_A^{(\mathcal{T}_1)^*}(f,t)\}, \end{aligned} \tag{7}$$

where $E\{\cdot\}$ denotes the expectation which can be obtained by averaging across the time frames, and $[\cdot]^*$ denotes the conjugate transpose. With further simplifications as in [15], we approximate

$$\boldsymbol{\mathcal{R}}_{AB}^{(N)}(f) \approx \boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)}(f)\left(\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_1)}(f)\right)^{\dagger}. \tag{8}$$

For convenience, we omit the dependency of time (t) and frequency (f) in the rest of this section.

B. Successive Speaker ReTF using Noise ReTM

Following the ReTM, our proposed ReTF estimation method is also based on covariance matrices between two microphone groups. In the second time segment, $\mathcal{T}_2$, the received signals at microphone groups $\{A\}$ and $\{B\}$ are given as

$$\mathbf{y}_A^{(\mathcal{T}_2)} = [\mathbf{h}_{A1}\ \mathbf{H}_{AN}][S_1\ \mathbf{N}]^T, \tag{9}$$

$$\mathbf{y}_B^{(\mathcal{T}_2)} = [\mathbf{h}_{B1}\ \mathbf{H}_{BN}][S_1\ \mathbf{N}]^T. \tag{10}$$

When all sources are active in the third time segment, $\mathcal{T}_3$, we obtain the received signals at microphone groups $\{A\}$ and $\{B\}$ as

$$\mathbf{y}_A^{(\mathcal{T}_3)} = [\mathbf{h}_{A1}\ \mathbf{h}_{A2}\ \mathbf{H}_{AN}][S_1\ S_2\ \mathbf{N}]^T, \tag{11}$$

$$\mathbf{y}_B^{(\mathcal{T}_3)} = [\mathbf{h}_{B1}\ \mathbf{h}_{B2}\ \mathbf{H}_{BN}][S_1\ S_2\ \mathbf{N}]^T. \tag{12}$$

Assuming all signal components in both (11) and (12) to be uncorrelated, the noisy covariance matrices for time segment $\mathcal{T}_3$ of microphone groups $\{A\}$ and $\{B\}$, given as

$$\begin{aligned} &\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_3)} \triangleq E\{\mathbf{y}_A^{(\mathcal{T}_3)}\mathbf{y}_A^{(\mathcal{T}_3)^H}\} \\ &= [\mathbf{h}_{A1}\ \mathbf{h}_{A2}\ \mathbf{H}_{AN}] \\ &\begin{bmatrix} E\{|S_1|^2\} & 0 & \mathbf{0}_{Q_A\times Q_A} \\ 0 & E\{|S_2|^2\} & \mathbf{0}_{Q_A\times Q_A} \\ \mathbf{0}_{Q_A\times Q_A} & \mathbf{0}_{Q_A\times Q_A} & \boldsymbol{\mathcal{P}}_{NN} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{A1}^* \\ \mathbf{h}_{A2}^* \\ \mathbf{H}_{AN}^H \end{bmatrix} \\ &= \mathbf{h}_{A1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{h}_{A2}E\{|S_2|^2\}\mathbf{h}_{A2}^* + \mathbf{H}_{AN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^* \end{aligned} \tag{13}$$

where $\boldsymbol{\mathcal{P}}_{NN} \triangleq E\{\mathbf{N}\mathbf{N}^*\}$.

Similarly, we can write

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_3)} = \mathbf{h}_{B1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{h}_{B2}E\{|S_2|^2\}\mathbf{h}_{A2}^* \\ + \mathbf{H}_{BN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*. \end{aligned} \tag{14}$$

Likewise, the covariance matrices for the time segment $\mathcal{T}_2$ of microphone groups $\{A\}$ and $\{B\}$, can be written as

$$\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_2)} = \mathbf{h}_{A1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{H}_{AN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*, \tag{15}$$

$$\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_2)} = \mathbf{h}_{B1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{H}_{BN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*. \tag{16}$$

We subtract (15) from (13) to remove both the first speaker and the background noise sources from the noisy mixture at group $\{A\}$. We then express the covariance matrix of the second speaker at group $\{A\}$ as

$$\begin{aligned} &\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_3)} - \boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_2)} = \mathbf{h}_{A2}E\{|S_2|^2\}\mathbf{h}_{A2}^*, \\ &= E\{|S_2|^2\}\mathbf{h}_{A2}\mathbf{h}_{A2}^*, \\ &= E\{|S_2|^2\}\underbrace{\begin{bmatrix} h_{A2}^{(1)}h_{A2}^{(1)*} & \dots & h_{A2}^{(1)}h_{A2}^{(Q_A)*} \\ \vdots & \ddots & \vdots \\ h_{A2}^{(Q_A)}h_{A2}^{(Q_A)*} & \dots & h_{A2}^{(Q_A)}h_{A2}^{(Q_A)*} \end{bmatrix}}_{\mathbf{R}_{AA}}, \end{aligned} \tag{17}$$

where $\mathbf{R}_{AA}$ is a $[Q_A \times Q_A]$ matrix of transfer function coefficients of the second speaker between channels at microphone group $\{A\}$.

Similarly, we subtract (16) from (14) as

$$\begin{aligned} &\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_3)} - \boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_2)} = E\{|S_2|^2\}\mathbf{h}_{B2}\mathbf{h}_{A2}^*, \\ &= E\{|S_2|^2\}\underbrace{\begin{bmatrix} h_{B2}^{(1)}h_{A2}^{(1)*} & \dots & h_{B2}^{(1)}h_{A2}^{(Q_A)*} \\ \vdots & \ddots & \vdots \\ h_{B2}^{(Q_A)}h_{A2}^{(Q_A)*} & \dots & h_{B2}^{(Q_A)}h_{A2}^{(Q_A)*} \end{bmatrix}}_{\mathbf{R}_{BA}}, \end{aligned} \tag{18}$$

where $\mathbf{R}_{BA}$ is a $[Q_B \times Q_A]$ matrix of transfer function coefficients of the second speaker between microphone groups $\{B\}$ and $\{A\}$.

Consider microphone channels n, and m at microphone group $\{A\}$ and $\{B\}$, respectively. We propose to divide $\{n,m\}$ elements of (17) by (18) to obtain ReTF $\{n,m\}$ pair with respect to the second speaker as

$$\begin{aligned} \hat{h}_2^{\{n,m\}} &= \frac{E\{|S_2|^2\}h_{A2}^{(n)}h_{A2}^{(n)*}}{E\{|S_2|^2\}h_{B2}^{(m)}h_{A2}^{(n)*}} \\ &= \frac{h_{A2}^{(n)}}{h_{B2}^{(m)}}. \end{aligned} \tag{19}$$

From (19), we obtain ReTF vector of the second speaker $\hat{\boldsymbol{h}}_2$ for $n \in 1,\dots,Q_A$. Therefore, the ReTF vector of the second speaker with respect to group $\{A\}$ as the reference channel at microphone group $\{B\}$ can be extracted from the covariance matrix addition.

Note that similarly, the ReTF of the first speaker, $\boldsymbol{h}_1$ can be obtained considering the covariance matrices at the time segments of $\mathcal{T}_1$ and $\mathcal{T}_2$ as in (17), and (18) and then followed by covariance matrices division in (19). However, in this paper, we cover neither the method nor the results of the ReTF of the first speaker from the noisy mixture.

IV. Experiments

This section presents experimental results for the proposed ReTF estimation algorithm using simulated recordings compared to the ground truth ReTFs.

A. Experiment Methodology

We utilize an open-source toolbox [25] to model the room impulse response (RIR) from the sound sources to irregularly distributed microphones in a $6 \times 7 \times 3$ m rectangular room with reverberation coefficients (β) of $\{0, 0.8, 1\}$. The ideal room with no reflection (free-space) would have $\beta = 0$, and $\beta = 1$ for the total reflection. We note that for a traditional implementation of a real room, the reverberation coefficients should be $0 < \beta < 1$. We consider two speech sources, two background noise sources, and 15 microphones. Two speaker locations are: Speaker 1: (3 m, 4.5 m, 1.2 m), Speaker 2: (3 m, 2.5 m, 1.2 m) and two background noise sources locations are: Source 1: (2 m, 0.9 m, 1.8 m), Source 2: (1 m, 6 m, 2.5 m) with respect to the origin position in the left corner of the room. We convolve the speech sources RIRs with both male (Speaker 1) and female (Speaker 2) speech utterances from the TIMIT dataset [26] and noise sources RIRs with vacuum noise, and music signal. The received signals are down-sampled to 16 kHz and ranged from 5 to -10 dB SNR of background noise and added with 40 dB SNR of white Gaussian noise at each microphone. Here, we calculate the background SNR by averaging SNR at each receiver over all 15 receivers. Note that the SNR is defined with respect to all sources in the mixture. The recordings are short-time-Fourier-transformed with an 8192-point window size. We assign $Q_A = 5$ and $Q_B = 10$ number of receivers to groups $\{A\}$ and $\{B\}$, respectively.

We use a distance measure based on the Hermitian angle [9] between the ground truth ReTF vector $\boldsymbol{h}_2$ and the estimated ReTF vector $\hat{\boldsymbol{h}}_2$ of the successive speaker as

$$\Theta(\mathbf{h}_2, \hat{\boldsymbol{h}}_2) = \arccos\Big(\frac{|\mathbf{h}_2^* \hat{\boldsymbol{h}}_2|}{||\mathbf{h}_2||_2 ||\hat{\boldsymbol{h}}_2||_2}\Big),$$

where $\arccos(\cdot)$ denotes inverse of the cosine function, $|\cdot|$ denotes the absolute value, and $||\cdot||_2$ denotes the ℓ_2 norm.

B. Results and Discussion

Fig. 2 shows the real and imaginary parts of the ground truth and estimated ReTFs of the successive speaker with respect to the second, and the first channels in the groups $\{A\}$ and $\{B\}$, respectively, for $\beta = 1$ at SNR level of 0 dB. We observe that many ReTF coefficients of both ground truth $\mathbf{h}_2$ and estimated $\hat{\boldsymbol{h}}_2$ are almost the same over the frequency up to 8 kHz. We discuss the results of these similarities between the ReTF vectors using the Hermitian angle in Table I.

Table I provides the distance measure results. We first observe that the proposed successive speaker ReTF estimation method is seen to have a lower Hermitian angle with the increase of the SNR level for all three reverberation coefficients. This suggests that the proposed method performs better at high SNR levels.

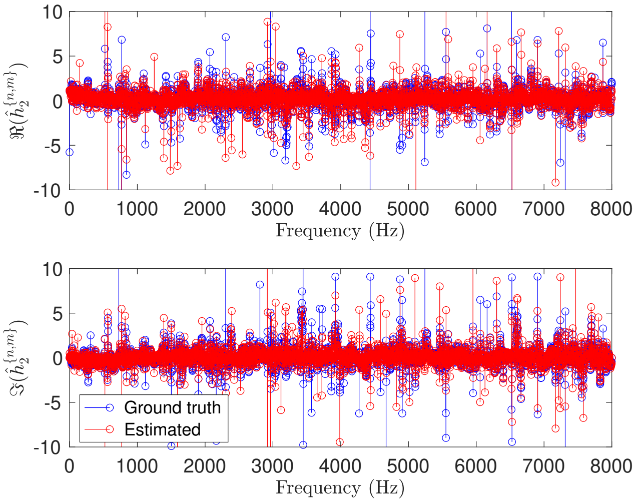

Fig. 2. The real and imaginary parts of the successive speaker ReTF obtained using the proposed method at microphone channel $n = 2, m = 1$ over the frequency at SNR level of 0 dB for $\beta = 1$.

TABLE I
DISTANCE MEASURE: HERMITIAN ANGLE FOR VARIOUS SNR LEVELS

SNR level	Reverberation coefficients		
	$\beta = 0$	$\beta = 0.8$	$\beta = 1$
5 dB	6.75°	8.95°	9.21°
0 dB	8.06°	11.94°	13.35°
-5 dB	10.55°	17.65°	17.86°
-10 dB	16.86°	27.68°	26.33°

Second, we observe that the higher distance measure results in increased reverberation coefficients. Further, the Hermitian angle had a lower variation in the angles for all three reverberation coefficients, however, is seen to obtain higher variation for the lower SNR conditions. This suggests that the proposed method is able to accurately estimate the ReTF of the successive speaker from the noisy multichannel recording of two speakers' mixture.

V. APPLICATIONS IN SPEECH ENHANCEMENT

In this section, we employ the minimum variance distortionless response (MVDR) beamformer for noise suppression and examine the performance in terms of the SNR improvement to evaluate the estimated ReTF vectors of the successive speaker in Section IV.

The MVDR beamformer requires an estimate of the noise-only covariance matrix that is equivalent to the received signals at $\mathcal{T}_1$ time segment in (7), and the estimate of the ReTFs of the second speaker, $\hat{\boldsymbol{h}}_2$. The weights of the MVDR beamformer can be expressed as in [27]

$$\mathbf{w} = \frac{{\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)}}^{-1} \hat{\boldsymbol{h}}_2}{\hat{\boldsymbol{h}}_2^* {\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)}}^{-1} \hat{\boldsymbol{h}}_2}, \tag{20}$$

where $(\cdot)^{-1}$ denotes the inversion operator, and $\mathbf{w}$ is the Q_A by 1 filter vector for a given frequency bin. The beamformer output is given by [28]

$$z(f, t) = \mathbf{w}^*(f) \mathbf{y}_A(f, t). \tag{21}$$

TABLE II
SNR IMPROVEMENT, ΔSNR FOR VARIOUS SNR LEVELS

SNR level	Reverberation coefficients		
	$\beta = 0$	$\beta = 0.8$	$\beta = 1$
5 dB	15.28 dB	12.16 dB	12.51 dB
0 dB	15.02 dB	15.42 dB	16.46 dB
-5 dB	14.46 dB	18.72 dB	18.88 dB
-10 dB	14.57 dB	21.64 dB	21.55 dB

We define the SNR improvement between the input and output of the beamformer as

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in,avg}},$$

where SNR_{out} is the filtered output SNR, and $\text{SNR}_{\text{in,avg}}$ is the average input SNR among all microphones in group $\{A\}$.

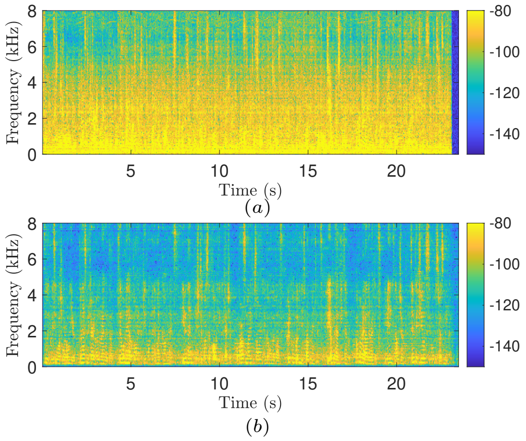

Fig. 3. Spectrogram plots of (a) microphone recordings at microphone channel 1 of Q_A, and (b) the extracted speech signals of the second speaker at 0 dB SNR level for $\beta = 0.8$ ($T_{60} = 500$ ms).

The spectrogram plots of the mixture and desired speech signals at the SNR level of 0 dB for $\beta = 0.8$ are shown in Fig. 3. Immediately we observe that the noise is removed and the voice of speaker 2 is shown to be successfully extracted at the beamformer output. We share the audio files of both input and output at the beamformer for $\beta = 0.8$ at all SNR levels[1].

Table II depicts the SNR improvement at the MVDR beamformer for the proposed ReTF estimation method as a function of the reverberation coefficient for $\beta = \{0, 0.8, 1\}$ over various SNR levels. We observe that in both reverberant environments, the SNR improvements are gradually increasing with decreased SNR levels. The SNR improvement at the beamformer obtains almost the same value with increased reverberation. A similar result with slight degradation is obtained for free-field scenario with an increased SNR level. As expected, the beamformer's output derived using the proposed ReTF estimation method is seen to be enhanced the successive speaker from the noisy mixture, illustrated by the high SNR improvement of 12 dB or above in both free-space and reverberant environments. The results confirm that the proposed ReTF method accurately enhances the voice of the second speaker in dual-speaker noisy reverberant rooms.

VI. Conclusion

We have proposed a ReTF estimation method for the second speaker in noisy and reverberant environments with two speech sources and multiple noise sources where the speech sources activate successively. The method utilizes the properties of the covariance matrices between two multichannel microphone groups imposed by the ReTM. Simulation analysis showed that the proposed method achieves a low Hermitian angle in both free-space and reverberant environments with differing SNR level. The MVDR beamformer which derives via the estimated ReTFs, is seen to have enhanced the voice of the successive speaker from the noisy microphone recordings with decreasing SNR levels. The accuracy of the ReTF estimation was increased within low reverberant environments, but SNR improvement was shown to improve with increasing reflection coefficient. In the future, we plan to extend this algorithm to the more challenging multiple-speaker scenarios exploiting the ReTM.

References

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[2] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 4, pp. 546–555, Mar. 2009.

[3] L. Birnie, P. Samarasinghe, T. Abhayapala, and D. Grixti-Cheng, "Noise retf estimation and removal for low snr speech enhancement," in *'IEEE Workshop Mach. Learning Signal Process.'*, Oct. 2021, pp. 1–6.

[4] W. Manamperi, P. N. Samarasinghe, T. D. Abhayapala, and J. Zhang, "GMM based multi-stage Wiener filtering for low SNR speech enhancement," in *Proc. IEEE Int. Workshop on Acoust. Signal Enhancement*, Sep. 2022, pp. 1–5.

[5] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2008, pp. 73–76.

[6] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. on Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.

[7] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 451–459, Aug. 2004.

[1] https://github.com/wnilmini/Successive_Speaker_ReTF_Estimation

[8] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, “Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 785–799, Feb. 2014.

[9] R. Varzandeh, M. Taseska, and E. A. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation,” in *Proc. Joint Workshop Hands-free Speech Comm. and Microphone Arrays*, Mar. 2017, pp. 11–15.

[10] S. Markovich, S. Gannot, and I. Cohen, “Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals,” *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Jun. 2009.

[11] E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jun. 2007.

[12] S. Markovich-Golan, S. Gannot, and W. Kellermann, “Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function,” in *Proc. Eur. Signal Process. Conf.*, Sep. 2018, pp. 2499–2503.

[13] A. Brendel, J. Zeitler, and W. Kellermann, “Manifold learning-supported estimation of relative transfer functions for spatial filtering,” in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 8792–8796.

[14] A. Deleforge, S. Gannot, and W. Kellermann, “Towards a generalization of relative transfer functions to more than one source,” in *Proc. Eur. Signal Process. Conf.*, Aug. 2015, pp. 419–423.

[15] T. D. Abhayapala, L. Birnie, M. Kumar, D. Grixti-Cheng, and P. N. Samarasinghe, “Generalizing the relative transfer function to a matrix for multiple sources and multichannel microphones,” in *Proc. Eur. Signal Process. Conf.*, Sep. 2023, pp. 336–340.

[16] M. Kumar and et al., “Speech denoising in multi-noise source environments using multiple microphone devices via relative transfer matrix,” in *Eur. Signal Process. Conf. (in press)*, Sep. 2024, pp. 336–340.

[17] W. N. Manamperi and T. D. Abhayapala, “Relative transfer matrix for drone audition applications: Source enhancement,” in *Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf. (in press)*, Dec. 2024.

[18] W. N. Manamperi and T. D. Abhayapala, “Blind separation of multiple speakers in noisy reverberant environments using relative transfer matrix,” in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (unpublished)*, 2024.

[19] B. Schwartz, S. Gannot, and E. A. Habets, “Two model-based EM algorithms for blind source separation in noisy environments,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 11, pp. 2209–2222, Aug. 2017.

[20] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Robust joint estimation of multimicrophone signal model parameters,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, Apr. 2019.

[21] C. Li, J. Martinez, and R. C. Hendriks, “Low complex accurate multi-source RTF estimation,” in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 4953–4957.

[22] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, “Square root-based multi-source early psd estimation and recursive RETF update in reverberant environments by means of the orthogonal procrustes problem,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 755–769, Jan. 2020.

[23] D. Cherkassky and S. Gannot, “Successive relative transfer function identification using blind oblique projection,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 474–486, Dec. 2019.

[24] H. Gode and S. Doclo, “Covariance blocking and whitening method for successive relative transfer function vector estimation in multi-speaker scenarios,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2023, pp. 1–5.

[25] E. A. Habets, *Room impulse response (RIR) generator*, [Online]. Available: https://www.audiolabserlangen.de/fau/professor/habets/software/rir-generator, 2006.

[26] J. S. Garofolo, “Timit acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium*, 1993.

[27] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[28] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 692–730, Jan. 2017.