

# RELATIVE TRANSFER MATRIX ESTIMATOR USING COVARIANCE SUBTRACTION

Wageesha N. Manamperi<sup>†,\*</sup>, and Thushara D. Abhayapala<sup>\*</sup>

<sup>†</sup> University of Moratuwa, Colombo, Sri Lanka

<sup>\*</sup> Australian National University, Canberra, Australia

## ABSTRACT

The Relative Transfer Matrix (ReTM), recently introduced as a generalization of the relative transfer function for multiple receivers and sources, shows promising performance when applied to speech enhancement and speaker separation in noisy environments. Blindly estimating the ReTM of sound sources by exploiting the covariance matrices of multichannel recordings is highly beneficial for practical applications. In this paper, we use covariance subtraction to present a flexible and practically viable method for estimating the ReTM for a select set of independent sound sources. To show the versatility of the method, we validated it through a speaker separation application under reverberant conditions. Separation performance is evaluated at low signal-to-noise ratio levels ( $\leq 0$  dB) in comparison with existing ReTM-based and relative transfer function-based estimators, in both simulated and real-life environments.

**Index Terms**— Covariance subtraction, multiple sound sources, relative transfer matrix, source separation

## 1. INTRODUCTION

Blindly estimating the relative transfer function (ReTF) without positional knowledge of receivers and sound source is highly attractive in practical applications such as robot audition, drone audition, teleconference, and hearing aids involving sound source localization, speech enhancement, speaker separation, acoustic echo cancellation [1–4]. Additionally, multiple simultaneously active sound sources are common in these applications and fail to hold W-disjoint orthogonality (WDO) [5] for obtaining ReTFs of sound sources in mixtures. As a result, the generalization of the ReTF for multiple sound sources introduced as the relative transfer matrix (ReTM) [6] is significantly beneficial in the field of audio signal processing. This paper presents a method to estimate the ReTM for a selected combination of active sources using the covariance subtraction property of independent signals.

Many approaches to estimate ReTF for a single active sound source have been developed [7–12]. Among them, covariance-based methods have gained attention due to their practical feasibility as well as simplicity [9–12]. ReTF for a multiple and concurrent source scenario has been proposed [13–17]. Common drawbacks to these approaches are

assuming WDO conditions, and a sufficiently high signal-to-background noise ratio (SNR).

Recently, two techniques have been proposed to generalize the ReTF for multiple simultaneous sound sources [6, 18]. In [18], the ReTF generalization was proposed to localize multiple sources, where the number of active sources and (either full or partially) activation of all sources were assumed for estimation.

In [6], the ReTM was introduced, which does not require source counting for speech enhancement or speaker separation in multi-source noisy reverberant environments, and subsequent works [19–22] have demonstrated its promising performance. Similar to ReTF, ReTM is independent of the signals emitted by multiple sources. By dividing multiple microphones into two groups, we can blindly estimate the ReTM from the received signals using the covariance-based method [6]. However, in practice, estimating the ReTM from segments where all desired sources are inactive in low-SNR scenarios is challenging [22]. Alternatively, this paper proposes a practically viable ReTM estimator for undesired sound sources using covariance subtraction [9–11] which estimates the ReTM with respect to a select group of sources to handle overlapping and continuously active sources. The proposed method’s applicability for speaker separation in a typical use case of a meeting scenario, is evaluated by referring to our previous work [22] and compared with the ReTF-based estimator using extensive experimental results in terms of speech intelligibility, signal-to-interference ratio, and signal-to-distortion ratio at very low SNR levels.

## 2. THE RELATIVE TRANSFER MATRIX

### 2.1. Signal model

Let us consider a reverberant environment with  $\mathcal{L}$  sound sources (both speech ( $S$ ) and background noise ( $N$ ) sources). In the short time Fourier transform (STFT) domain, we denote  $S_\ell(f, t)$ ,  $\ell = 1, \dots, \mathcal{L}$  as the source signal.

Let there be  $Q$  arbitrary distributed microphones in the room. We divide them to two groups of microphones,  $\{A\}$  and  $\{B\}$  with  $Q_A$  and  $Q_B$  microphones, respectively ( $Q = Q_A + Q_B$ ). We denote  $\mathbf{M}_A(f, t)$  and  $\mathbf{M}_B(f, t)$  as the vector of received signals at microphone groups A and B, respectively. Then the received signals at each microphone group in

matrix form as

$$\mathbf{M}_A(f, t) = \mathbf{H}_A(f) \mathbf{S}(f, t), \text{ and} \quad (1)$$

$$\mathbf{M}_B(f, t) = \mathbf{H}_B(f) \mathbf{S}(f, t), \text{ where} \quad (2)$$

$$\mathbf{S}(f, t) = [S_1, \dots, S_\ell, \dots, S_{\mathcal{L}}]^T = [\mathbf{S}^{(S)}(f, t) \quad \mathbf{S}^{(N)}(f, t)]^T$$

and  $\{\cdot\}^T$  is the matrix transpose. Note that  $\mathbf{S}^{(S)}$  and  $\mathbf{S}^{(N)}$  are the vectors of speech and background noise source signals, respectively. Also,  $\mathbf{H}_A(f) \in \mathbb{C}^{Q_A \times \mathcal{L}}$  and  $\mathbf{H}_B(f) \in \mathbb{C}^{Q_B \times \mathcal{L}}$  are the matrices with elements defined by the acoustic transfer functions, and defined as  $\mathbf{H}_A(f) = [\mathbf{H}_A^{(S)}(f) \quad \mathbf{H}_A^{(N)}(f)]$ , and  $\mathbf{H}_B(f) = [\mathbf{H}_B^{(S)}(f) \quad \mathbf{H}_B^{(N)}(f)]$ .

The relative transfer matrix (ReTM) with respect to all active sources,  $\mathcal{R}_{AB}(f)$ , is defined as in [6]

$$\mathcal{R}_{AB}(f) \triangleq \mathbf{H}_A(f) \mathbf{H}_B(f)^\dagger, \quad (3)$$

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

$$\mathbf{M}_A(f, t) = \mathcal{R}_{AB}(f) \mathbf{M}_B(f, t).$$

Note that  $\mathcal{R}_{AB}(f)$  is defined by the spatial properties of the sound sources and the acoustic environment such that it is independent of the sound source signals. In applications, the ReTM is invariant for a wide-sense stationary (WSS) acoustic environment.

## 2.2. ReTM estimation with covariance matrices

In practice, the ReTM of the sound sources can be estimated using a segment of the microphone recording using the covariance matrices-based approach [6], given as

$$\mathcal{R}_{AB}(f) \approx \mathcal{P}_{AA}(f) \left( \mathcal{P}_{BA}(f) \right)^\dagger, \text{ where} \quad (4)$$

$$\mathcal{P}_{AA}(f) \triangleq E\{\mathbf{M}_A(f, t) \mathbf{M}_A^*(f, t)\}, \quad (5)$$

$$\mathcal{P}_{BA}(f) \triangleq E\{\mathbf{M}_B(f, t) \mathbf{M}_A^*(f, t)\}, \quad (6)$$

$E\{\cdot\}$  denotes the expectation which can be obtained by averaging across the time frames, and  $[\cdot]^*$  denotes the conjugate transpose. To make the exposition concise, we omit the dependency of time ( $t$ ) and frequency ( $f$ ) in the following sections. This paper aims to develop a practical ReTM estimator that does not require the segments of a set of sources to be inactive, with covariance subtraction, which we propose next.

## 3. PROPOSED PRACTICAL RETM ESTIMATOR USING COVARIANCE SUBTRACTION

This section uses covariance matrices to propose a method for estimating the ReTM of a selected set of individual sound sources using covariance subtraction.

Expanding (5) by substituting (1) as

$$\begin{aligned} \mathcal{P}_{AA} &= E\left\{ \begin{bmatrix} \mathbf{H}_A^{(S)} & \mathbf{H}_A^{(N)} \end{bmatrix} \begin{bmatrix} \mathbf{S}^{(S)} \\ \mathbf{S}^{(N)} \end{bmatrix} \begin{bmatrix} \mathbf{S}^{(S)*} & \mathbf{S}^{(N)*} \end{bmatrix} \begin{bmatrix} \mathbf{H}_A^{(S)*} \\ \mathbf{H}_A^{(N)*} \end{bmatrix} \right\}, \\ &= \begin{bmatrix} \mathbf{H}_A^{(S)} & \mathbf{H}_A^{(N)} \end{bmatrix} \begin{bmatrix} \mathcal{P}_{SS} & \mathbf{0}_{Q_A \times Q_A} \\ \mathbf{0}_{Q_A \times Q_A} & \mathcal{P}_{NN} \end{bmatrix} \begin{bmatrix} \mathbf{H}_A^{(S)*} \\ \mathbf{H}_A^{(N)*} \end{bmatrix}, \\ &= \begin{bmatrix} \mathbf{H}_A^{(S)} \mathcal{P}_{SS} & \mathbf{H}_A^{(N)} \mathcal{P}_{NN} \end{bmatrix} \begin{bmatrix} \mathbf{H}_A^{(S)*} \\ \mathbf{H}_A^{(N)*} \end{bmatrix}, \\ &= \underbrace{\mathbf{H}_A^{(S)} \mathcal{P}_{SS} \mathbf{H}_A^{(S)*}}_{\mathcal{P}_{AA}^{(S)}} + \underbrace{\mathbf{H}_A^{(N)} \mathcal{P}_{NN} \mathbf{H}_A^{(N)*}}_{\mathcal{P}_{AA}^{(N)}}. \end{aligned} \quad (7)$$

where  $\mathcal{P}_{SS} = E\{\mathbf{S}^{(S)} \mathbf{S}^{(S)*}\}$ , and  $\mathcal{P}_{NN} = E\{\mathbf{S}^{(N)} \mathbf{S}^{(N)*}\}$  are the auto-covariance matrices of the speech and noise source signals, respectively.

Similarly, substituting (2) into (6), we can write

$$\begin{aligned} \mathcal{P}_{BA} &= \begin{bmatrix} \mathbf{H}_B^{(S)} \mathcal{P}_{SS} & \mathbf{H}_B^{(N)} \mathcal{P}_{NN} \end{bmatrix} \begin{bmatrix} \mathbf{H}_A^{(S)*} \\ \mathbf{H}_A^{(N)*} \end{bmatrix}, \\ &= \underbrace{\mathbf{H}_B^{(S)} \mathcal{P}_{SS} \mathbf{H}_A^{(S)*}}_{\mathcal{P}_{BA}^{(S)}} + \underbrace{\mathbf{H}_B^{(N)} \mathcal{P}_{NN} \mathbf{H}_A^{(N)*}}_{\mathcal{P}_{BA}^{(N)}}. \end{aligned} \quad (8)$$

Rearranging (7) gives an expression for the speech sources' covariance matrix of microphone group  $\{A\}$  in terms of the covariance matrices of the received signal and the noise source component of the received signal,

$$\mathcal{P}_{AA}^{(S)} = \mathcal{P}_{AA} - \mathcal{P}_{AA}^{(N)}. \quad (9)$$

From (8), the speech sources' covariance matrix of microphone groups  $\{A\}$  and  $\{B\}$  can be estimated as

$$\mathcal{P}_{BA}^{(S)} = \mathcal{P}_{BA} - \mathcal{P}_{BA}^{(N)}. \quad (10)$$

Let  $\mathcal{R}_{AB}^{(S)}$  be the ReTM of the speech sources. Substituting (9), and (10) back into (4) provides the ReTM of speech sources as

$$\mathcal{R}_{AB}^{(S)} = \left( \mathcal{P}_{AA} - \mathcal{P}_{AA}^{(N)} \right) \left( \mathcal{P}_{BA} - \mathcal{P}_{BA}^{(N)} \right)^\dagger. \quad (11)$$

We can observe from (11) that the ReTM of the speech sources can be estimated by using the covariance matrices of the received signals and the received noise-only signals. This allows us to confirm the ReTM of speech sources can be estimated using covariance subtraction.

We note that the ReTM of the background noise sources  $\mathcal{R}_{AB}^{(N)}$  can be blindly estimated from the received noise-only signals (no active speech) as

$$\mathcal{R}_{AB}^{(N)} = \mathcal{P}_{AA}^{(N)} \left( \mathcal{P}_{BA}^{(N)} \right)^\dagger. \quad (12)$$

Furthermore, rearranging (4) (by substituting both (7) and (8)) gives an expression for the ReTM of all sound sources

in terms of the covariance matrices of the speech and noise sources between microphone groups  $\{A\}$  and  $\{B\}$  as

$$\mathcal{R}_{AB} = (\mathcal{P}_{AA}^{(S)} + \mathcal{P}_{AA}^{(N)}) (\mathcal{P}_{BA}^{(S)} + \mathcal{P}_{BA}^{(N)})^\dagger. \quad (13)$$

We can observe from (13) that, unlike the covariance matrices of the speech and noise sources, the ReTM of the speech and noise sources cannot be added together. Hence, ReTMs are not additive. However, we observe that covariance matrices can be manipulated to derive the ReTM of the speech sources. This property then allows for the generalization of ReTM estimation to be calculated by a select set of independent sound sources.

Following the simplifications in (7) and (8), we can decompose the covariance matrices of independent  $\mathcal{L}$  sources between microphone groups  $\{A\}$  and  $\{B\}$  into its individual source covariance matrices,  $\mathcal{P}_{AA}^{(\ell)}$  and  $\mathcal{P}_{BA}^{(\ell)}$ , expressed as

$$\mathcal{P}_{AA} = \sum_{\ell=1}^{\mathcal{L}} \mathbf{h}_A^{(\ell)} \mathcal{P}_\ell \mathbf{h}_A^{(\ell)*} = \sum_{\ell=1}^{\mathcal{L}} \mathcal{P}_{AA}^{(\ell)}, \quad (14)$$

$$\mathcal{P}_{BA} = \sum_{\ell=1}^{\mathcal{L}} \mathbf{h}_B^{(\ell)} \mathcal{P}_\ell \mathbf{h}_A^{(\ell)*} = \sum_{\ell=1}^{\mathcal{L}} \mathcal{P}_{BA}^{(\ell)}, \quad (15)$$

where  $\mathbf{h}_A^{(\ell)}$  and  $\mathbf{h}_B^{(\ell)}$  be the acoustic transfer function vectors from the  $\ell^{\text{th}}$  source to group  $\{A\}$  and  $\{B\}$  microphones, respectively, and  $\mathcal{P}_\ell \triangleq E\{|S_\ell|^2\}$ .

Therefore, following the covariance subtraction, a set of independent  $\hat{\mathcal{L}}$  sources can be used to practically generalize the ReTM, given as

$$\hat{\mathcal{R}}_{AB} = \left( \sum_{\ell=1}^{\hat{\mathcal{L}}} \mathcal{P}_{AA}^{(\ell)} \right) \left( \sum_{\ell=1}^{\hat{\mathcal{L}}} \mathcal{P}_{BA}^{(\ell)} \right)^\dagger. \quad (16)$$

Here, we note that a set of different active combinations of sources can be used to practically manipulate the ReTM estimation in (16) via (14) and (15).

The next section utilizes the proposed ReTM estimator to develop a more practically viable method for ReTM-based speaker separation.

#### 4. APPLICATION INTO SPEAKER SEPARATION

This section employs the proposed ReTM estimator to extract individual speakers from mixtures of multiple speech and noise sources, following the approach in [22].

Let us consider the number of concurrently active  $\mathcal{L}_S$  speech and  $\mathcal{L}_N$  background noise sources, where  $\mathcal{L} = \mathcal{L}_S + \mathcal{L}_N$ . In STFT domain, we denote  $S_\ell^{(S)}$ ,  $\ell = 1, \dots, \mathcal{L}_S$  and  $S_n^{(N)}$ ,  $n = 1, \dots, \mathcal{L}_N$  as the speech and background noise signals, respectively.

In the following, we extract the target speech of the  $\ell^{\text{th}}$  speaker. Note that all speakers,  $\ell = 1, \dots, \mathcal{L}_S$ , in the mixture can be similarly extracted. Let the target  $\ell^{\text{th}}$  speech denotes as

$S_\ell^{(S)}$  out of  $\mathcal{L}_S$  concurrent speakers. The rest of the undesired source signals including background noise can be grouped as  $\bar{\mathbf{S}}_\ell = [S_1^{(S)} \dots S_{\ell-1}^{(S)}, S_{\ell+1}^{(S)} \dots S_{\mathcal{L}_S}^{(S)}, S_1^{(N)} \dots S_{\mathcal{L}_N}^{(N)}]^T$ . Here we redefine the source signal vector as  $\mathbf{S}(f, t) = [S_\ell^{(S)} \bar{\mathbf{S}}_\ell^T]^T$ .

Let  $\bar{\mathbf{H}}_A^{(\ell)}$  and  $\bar{\mathbf{H}}_B^{(\ell)}$  be the acoustic transfer function matrices from all other sources except the  $\ell^{\text{th}}$  speech source to group A and B microphones, respectively. Note that  $\mathbf{H}_A = [\mathbf{h}_A^{(\ell)} \bar{\mathbf{H}}_A^{(\ell)}]$  and  $\mathbf{H}_B = [\mathbf{h}_B^{(\ell)} \bar{\mathbf{H}}_B^{(\ell)}]$ .

Denote  $\bar{\mathcal{R}}_{AB}^{(\ell)}$  be the ReTM of the combination of all sound sources except the  $\ell^{\text{th}}$  target source. The separated speech signal of the  $\ell^{\text{th}}$  speaker is then given by [22]

$$\hat{\mathbf{S}}_\ell \triangleq \mathbf{M}_A - \bar{\mathcal{R}}_{AB}^{(\ell)} \mathbf{M}_B, \quad (17)$$

$$= [\mathbf{H}_A - \bar{\mathcal{R}}_{AB}^{(\ell)} \mathbf{H}_B] [S_\ell^{(S)} \bar{\mathbf{S}}_\ell^T]^T, \quad (18)$$

$$\begin{aligned} &= [\mathbf{h}_A^{(\ell)} - \bar{\mathcal{R}}_{AB}^{(\ell)} \mathbf{h}_B^{(\ell)}] S_\ell^{(S)} + \underbrace{[\bar{\mathbf{H}}_A^{(\ell)} - \bar{\mathcal{R}}_{AB}^{(\ell)} \bar{\mathbf{H}}_B^{(\ell)}] \bar{\mathbf{S}}_\ell}_{\approx \mathbf{0}} \\ &= \underbrace{[\mathbf{h}_A^{(\ell)} - \bar{\mathcal{R}}_{AB}^{(\ell)} \mathbf{h}_B^{(\ell)}]}_{\text{distortion}} S_\ell^{(S)}. \end{aligned} \quad (19)$$

We note that  $\hat{\mathbf{S}}_\ell$  is a  $Q_A \times 1$  vector consists  $Q_A$  copies of estimated target speech signal  $S_\ell^{(S)}$ .

In brief, [22] proposed to extract the desired speaker from (17) with known segment boundaries of the undesired sources-only signals from the microphone recording to calculate the  $\bar{\mathcal{R}}_{AB}^{(\ell)}$  as

$$\bar{\mathcal{R}}_{AB}^{(\ell)} \approx \bar{\mathcal{P}}_{AA}^{(\ell)} \left( \bar{\mathcal{P}}_{BA}^{(\ell)} \right)^\dagger. \quad (20)$$

Following the procedure in Section 3, i.e., adopting (9), (10), (11), and (16), we have (20) as

$$\begin{aligned} \bar{\mathcal{R}}_{AB}^{(\ell)} &= \left( \sum_{\substack{\ell'=0 \\ \ell' \neq \ell}}^{\mathcal{L}_S} (\mathcal{P}_{AA}^{(\mathcal{L}_N, \ell')} - \mathcal{P}_{AA}^{(\mathcal{L}_N)}) + \mathcal{P}_{AA}^{(\mathcal{L}_N)} \right) \times \\ &\quad \left( \sum_{\substack{\ell'=0 \\ \ell' \neq \ell}}^{\mathcal{L}_S} (\mathcal{P}_{BA}^{(\mathcal{L}_N, \ell')} - \mathcal{P}_{BA}^{(\mathcal{L}_N)}) + \mathcal{P}_{BA}^{(\mathcal{L}_N)} \right)^\dagger. \end{aligned} \quad (21)$$

where  $\mathcal{P}_{AA}^{(\mathcal{L}_N)}$  and  $\mathcal{P}_{AA}^{(\mathcal{L}_N, \ell')}$  are the covariance matrices of background noise-only signals, and background noise plus  $\ell'$  speaker for  $\ell' = 1, \dots, \mathcal{L}_S$ , respectively.

In teleconferencing, for a given meeting room with a fixed seating arrangement, both  $\mathcal{P}_{AA}^{(\mathcal{L}_N)}$  and  $\mathcal{P}_{AA}^{(\mathcal{L}_N, \ell')}$ ,  $\ell' = 1, \dots, \mathcal{L}_S$  can be recorded before the session begins. This algorithm does not require explicit pre-speech segments for each source, instead, it relies on the approximate independence of the speech sources, whereby their covariance contributions add up.

**Table 1.** Speaker separation results for various SNR levels (SIR (dB)/SDR (dB)/STOI (%)).

SNR level	Method (SIR (dB)/SDR (dB)/STOI (%))			
	Unprocessed	ReTF - Oracle	ReTM (eq. (4))	Proposed
−20 dB	−4.15/−21.14/29.66	27.29/ <b>4.84</b> /68.37	<b>30.37</b> /2.73/ <b>72.03</b>	30.36/2.73/72.02
−15 dB	−4.35/−17.69/31.21	27.54/ <b>4.83</b> /68.48	<b>30.52</b> /2.71/71.80	30.51/2.73/ <b>71.87</b>
−10 dB	−4.45/−13.59/33.95	27.63/ <b>4.82</b> /68.51	<b>30.38</b> /2.52/ <b>71.78</b>	30.35/2.50/71.77
−5 dB	−4.48/−9.74/37.86	27.66/ <b>4.78</b> /68.49	<b>30.31</b> /2.67/71.85	30.30/2.65/ <b>72.02</b>
0 dB	−4.48/−7.44/41.44	27.63/ <b>4.76</b> /68.45	30.43/3.15/72.17	<b>30.46</b> /3.13/ <b>72.60</b>

## 5. EXPERIMENTS

This section presents experimental results using the proposed ReTM estimator for speaker separation with both simulated and real-life recordings at low ( $\leq 0$  dB) signal-to-background noise ratio (SNR) levels.

We evaluate speaker separation performance using (i) Signal-to-Interference Ratio (SIR), (ii) Signal-to-Distortion Ratio (SDR) (in BSS-Eval toolbox [23]), and (iii) Short-Time Objective Intelligibility (STOI) [24]. The SNR is defined with respect to all sources in the mixture, whereas SIR is defined considering one speaker as the target signal and the rest of the interfering speakers as the noise signal. To facilitate the interpretation of the results, we assume the availability of the oracle ReTF with its MVDR beamformer estimate [20].

### 5.1. Simulated Environments

We utilize an open-source toolbox [25] to model the room impulse response (RIR) from sound sources to irregularly distributed microphones in a  $6 \times 7 \times 3$  m rectangular room ( $T_{60} = 500$  ms). We consider 3 speech sources, 2 background noise sources, and 27 microphones. We convolve the speech sources RIRs with both male and female speech utterances from the TIMIT dataset [26] and noise sources RIRs with wall air-conditioner noise, and vacuum noise. The received signals are ranged from 0 to −20 dB SNR of background noise and added with 40 dB SNR of white Gaussian noise at each microphone. Here, we define the SNR by averaging the SNR at each receiver over all 27 receivers. The short (60 second) recordings of background noise sources only, background noise sources plus each speaker are obtained for  $\bar{\mathcal{R}}_{AB}^{(\ell)}$  training from (21). These recordings are short-time-Fourier-transformed with an 8192-point window size that was long relative to the length of the RIR to satisfy the multiplicative transfer function [27]. We assign  $Q_A = 10$ , and  $Q_B = 17$  number of receivers to group  $\{A\}$  and  $\{B\}$ , respectively.

Table 1 depicts the speaker separation performance in various noisy environments. The results confirm that both ReTF- and ReTM-based estimators accurately separate all speakers in very low SNR levels. We observe the highest output SDR with the ReTF estimator, whereas the highest SIR and STOI values are achieved with the ReTM estimators, as the improved SIR leads to a slight reduction in SDR. However, we note that accurately estimating the ReTF is difficult in reverberant rooms with multiple noise sources. We also observe

that the ReTM from (4) and the proposed ReTM estimator exhibit a similar performance.

### 5.2. Real-life Environments

**Table 2.** Speaker separation results at low SNR =  $\{-10, 0\}$  dB for real recordings. The values are increments with respect to the unprocessed signals and averaged across speakers ( $\Delta$ SIR (dB)/ $\Delta$ SDR (dB)/ $\Delta$ STOI (%))

SNR level	Method ( $\Delta$ SIR (dB)/ $\Delta$ SDR (dB)/ $\Delta$ STOI (%))	
	ReTF - Oracle	Proposed
−10 dB	4.70/0.15/1.78	<b>11.42/8.13/19.93</b>
0 dB	20.15/1.4/3.46	<b>26.12/4.9/31.78</b>

Next, we examine the performance of the proposed method in real-life scenarios. The experimental recordings are measured in an office room at the Australian National University with dimensions  $2.95 \times 6 \times 3.03$  m, and  $T_{60} \approx 630$  ms. We consider 5 loudspeakers, including 3 speakers, 2 background noise sources (fan and room air cooler), and 15 randomly distributed microphones over the room. We assigned  $Q_A = 5$ , and  $Q_B = 10$  number of receivers to group  $\{A\}$  and  $\{B\}$ , respectively. We examined both the ‘Proposed’ and ‘ReTF - Oracle’ estimators to separation performance in Table 2. Again, the proposed method consistently improves the average SIR, SDR, and STOI over the mixture signals for all three speakers compared to the baseline. The ‘Proposed’ estimator performs as anticipated by simulation analysis conclusions, with slightly lower results due to unavoidable practical errors.

## 6. CONCLUSION

This paper proposed a novel method to estimate the ReTM using covariance matrices in a noisy, reverberant environment with multiple speech and noise sources. The method uses covariance subtraction to estimate the ReTM for different subgroups of sound sources. We showed that the ReTM is not necessarily additive. We generalized the ReTM estimation for a select set of independent sources to manipulate the covariance matrices between microphone groups along with covariance subtraction. We evaluated this method on a speaker separation application using both simulation and real-life recordings. Extensive experimental study confirmed that this method achieves strong separation results under very low SNR conditions. In the future, we plan to utilize this method to sound zone control problem.

## 7. REFERENCES

- [1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Natl. Conf. Artif. Intell.*, Jul. 2000, pp. 832–839.
- [2] W. N. Manamperi, T. D. Abhayapala, L. Brinie, J. Zhang, and P. N. Samarasinghe, "Drone audition: On measurements and modeling of drone-related transfer functions," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 33, pp. 1775–1786, Apr. 2025.
- [3] 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.
- [4] W. N. Manamperi, T. D. Abhayapala, P. N. Samarasinghe, and J. A. Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones using multichannel wiener filtering and gaussian-mixture based post-filtering," *Appl. Acoust.*, vol. 216, pp. 109818, Jan. 2024.
- [5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Signal Process. Mag.*, vol. 52, no. 7, pp. 1830–1847, Jun. 2004.
- [6] 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.
- [7] O. Shalvi and E. Weinstein, "System identification using non-stationary signals," *IEEE Trans. on Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.
- [8] 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.
- [9] W. Middelberg, H. Gode, and S. Doclo, "Relative transfer function vector estimation for acoustic sensor networks exploiting covariance matrix structure," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2023, pp. 1–5.
- [10] 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.
- [11] R. Varzandeh, M. Taseska, and E. A. P. 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.
- [12] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Jun. 2009.
- [13] B. Schwartz, S. Gannot, and E. A. P. 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] M. Kumar, L. Birnie, T. Abhayapala, S. A. Holzinger, A. Bastine, D. Grixti-Cheng, and P. Samarasinghe, "Speech denoising in multi-noise source environments using multiple microphone devices via relative transfer matrix," in *Proc. Eur. Signal Process. Conf.*, Sep. 2024, pp. 336–340.
- [20] W. N. Manamperi and T. D. Abhayapala, "Successive speaker relative transfer function estimation through relative transfer matrix in noisy reverberant environments," in *Proc. Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf.*, Dec. 2024, pp. 1–6.
- [21] W. N. Manamperi and T. D. Abhayapala, "Relative transfer matrix for drone audition applications: Source enhancement," in *Proc. Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf.*, Dec. 2024, pp. 1–6.
- [22] W. N. Manamperi and T. D. Abhayapala, "Multiple speaker separation from noisy sources in reverberant rooms using relative transfer matrix," *arXiv preprint arXiv:2503.09412*, Mar. 2025.
- [23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.
- [24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.
- [25] E. A. Habets, "Room impulse response (RIR) generator," 2006, [Online]. Available: <https://www.audiolabserlangen.de/fau/professor/habets/software/rir-generator>.
- [26] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.
- [27] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Letters*, vol. 14, no. 5, pp. 337–340, Apr. 2007.