GENERALIZING THE RELATIVE TRANSFER FUNCTION TO A MATRIX FOR MULTIPLE SOURCES AND MULTICHANNEL MICROPHONES

Thushara D. Abhayapala, Lachlan Birnie, Manish Kumar, Daniel Grixti-Cheng, Prasanga N. Samarasinghe

Audio and Acoustic Signal Processing Group, The Australian National University, Canberra, Australia

ABSTRACT

The Relative Transfer Function (ReTF) between two microphones with respect to a source location has become an integral feature in many acoustic signal processing and learning applications. However, when there are multiple simultaneously active sources and receivers, estimation of ReTF of each source is not possible. In this paper, we generalize the ReTF concept to multiple sources and receivers. We divide the multiple receivers into two groups and formulate the Relative Transfer Matrix (ReTM) to describe the coupling between them in response to multiple sound sources. We show that the ReTM is independent of source signals, captures spatial properties of sources, and can be directly estimated from the observed signals. We illustrate the validity of the concept through preliminary simulations and experimental recordings.

Index Terms— Relative Transfer Function, Relative Transfer Matrix, Spatial Audio, Microphone Array.

1. INTRODUCTION

Estimating the relative transfer functions (ReTFs) between two microphones when there are multiple sources present is an outstanding problem today. Conventionally, ReTF literature handles multiple sources by assuming that they alternate in activity during the estimation process [1]. This assumption, however, is often not feasible in practical applications. Especially when one sound source is continuously active. Finding a generalized ReTF solution for multiple sources has not progressed much, as a straightforward extension is difficult [2]. In this paper, we propose a systematic extension to the ReTF concept for multiple sound sources and microphones.

The ReTF is given by the ratio of acoustic transfer functions between a source and receivers [3, 4]. In a sense, it describes the acoustic channel or coupling between two microphones in response to a sound source. Furthermore, the ReTF has the three properties:

i) It is a unique signature of the source-microphone position and the environmental characteristics, such as room size and reverberation [3, 5].
ii) It is a spatial property that is independent of emitted signals.
iii) It can be reliably and robustly estimated directly from observed multichannel signals [5, 6, 7, 8]. Given that only one source is active during the estimation.

These three spatial properties of the ReTF make it an ideal tool for various signal processing applications, such as blind source separation [9, 10, 11], beamforming [6, 12, 13], sound source localization [3, 14, 15], acoustic echo cancellation [16], microphone array calibration [4], speech enhancement [17], and hearing aids [18]. In more recent years, the ReTF has also been used as a feature in learning algorithms for source localization [19] and low SNR (< -10 dB) speech enhancement [5]. The direct-path ReTF has also been defined and used in many applications [20]. Moreover, the ReTF concept has been extended to the spherical harmonic domain as Relative Harmonic Coefficients [21] and been used for multiple source localization [22, 23].

Deleforge et al. has shown that a natural generalization of the ReTF to more than one source is not possible using a single multichannel spectro-temporal observation [2] (which we review in Sec. 2). To combat this, they propose a generalized solution by introducing a transform containing multichannel multi-frame spectrograms. To the best of our knowledge, the work of Deleforge et al. is the only reported attempt of generalizing the ReTF.

In this paper, we extend the ReTF to the *Relative Transfer Matrix* (ReTM) for multiple sources and receivers, by considering two multichannel groups of microphones. Noting that the ReTF can be viewed as a spatial mapping between the received signal at one microphone to the received signal at a second microphone due to the single sound source, we consider a spatial mapping matrix between the received signal vector at the first group of microphones to the received signal vector at the second group of microphones due to multiple sources. This new concept still possesses three desirable properties similar to the ReTF, i.e., the ReTM is:

i) independent of the emitted source signals;
ii) a unique spatial property of the source-microphone position and the environment; and
iii) can be estimated directly from observed signals.

We derive the ReTM in Sec. 3, and provide a covariance based method to estimating the ReTM from observed signals in Sec. 3.1. Finally, in Sec. 4.1, we provide a preliminary validation of the ReTM through an example use case based on numerical simulation and live experimental recordings.

2. PROBLEM FORMULATION

In this section, we review the ReTF in theory and its estimation in practice. We then review the case where there are multiple simultaneously active sound sources.

Beforehand, let us define the acoustic system as illustrated in Fig. 1. Consider two microphones indexed by integers $\{1, 2\}$ and for now a single sound source denoted by $\{\alpha\}$. We will consider the second source, $\{\beta\}$, later on. Denote the acoustic transfer functions between the source and the microphones as $H_{1\alpha}$ and $H_{2\alpha}$. The short-time frequency domain signals received by the microphones, $M_{\{1,2\}}$, are described as

$$M_1(f,t) = H_{1\alpha}(f) S_\alpha(f,t), \quad \text{and} \quad M_2 = H_{2\alpha} S_\alpha, \tag{1}$$

where (f, t) denotes frequency and time, and S is the source's signal. Note that we often drop the (f, t) notation for brevity.

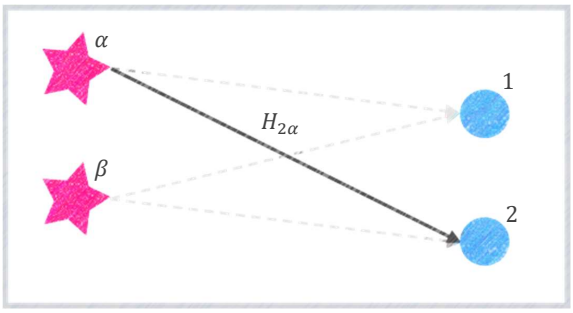

Fig. 1. Drawing of receivers, sources, and transfer functions.

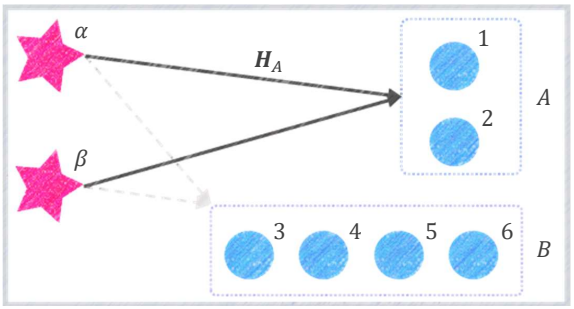

Fig. 2. Illustration of grouped microphones.

2.1. The relative transfer function

The ReTF describes the coupled response between two sound receivers due to the stimulus of a single sound source. It is a spatial function dependent on the environment, the position of the receivers, and the position of the sound source. Moreover, it is a function that is independent of the source's emitted signal content. Therefore, the ReTF is a constant spatial feature of an acoustic system when the environment, sources, and receivers are unmoving.

In theory, the ReTF is given by the ratio of the acoustic transfer functions between the source and two receivers. The ReTFs for the source α are given by

$$\mathcal{R}^{\alpha}_{1,2}(f) = H_{1\alpha}(f)/H_{2\alpha}(f), \quad \text{and} \quad \mathcal{R}^{\alpha}_{2,1} = H_{2\alpha}/H_{1\alpha}, \tag{2}$$

depending on whether microphone-$\{2\}$ or -$\{1\}$ is taken as the reference, respectively. The ReTF can be estimated using (2) when the acoustic transfer functions have been measured. This, however, typically requires impulse response (IR) measurements of the source and a stable acoustic environment, which is difficult in practice.

In another sense, the ReTF of $\mathcal{R}^{\alpha}_{1,2}$ can be thought to be the spatial mapping between the received signal at microphone-$\{2\}$ to the received signal at microphone-$\{1\}$, due to the single sound source $\{\alpha\}$. This mapping is illustrated by

$$M_1(f,t) \; = \mathcal{R}^{\alpha}_{1,2}(f) M_2(f,t) \; = H_{1\alpha}(f) S_{\alpha}(f,t). \tag{3}$$

2.2. Estimating the ReTF from received signals

The ReTF has a benefit in that it can be estimated directly from measured signals in a live environment. Consider the simple example of taking the ratio of the two received signals in (1), expressed as

$$\frac{M_1(f,t)}{M_2(f,t)} = \frac{H_{1\alpha}(f)S_{\alpha}(f,t)}{H_{2\alpha}(f)S_{\alpha}(f,t)} = \frac{H_{1\alpha}(f)}{H_{2\alpha}(f)} = \mathcal{R}^{\alpha}_{1,2}(f). \tag{4}$$

Observe that the source's signal elegantly cancels off in (4), leaving behind only spatial information about the source and environment. We note that the ratio of (4) is susceptible to noise in practice. Instead, it helps to estimate the ReTF with the cross power spectral density, $\mathcal{P}$, of the received signals, [24]

$$\mathcal{R}^{\alpha}_{1,2}(f) \approx \mathcal{P}_{1,2}(f,t)/\mathcal{P}_{2,2}(f,t). \tag{5}$$

2.3. The ReTF with multiple sources

Let us now consider a second sound source denoted by $\{\beta\}$. The received signals analogous to (1) are now given by

$$M_1 = H_{1\alpha}S_{\alpha} + H_{1\beta}S_{\beta}, \quad \text{and} \quad M_2 = H_{2\alpha}S_{\alpha} + H_{2\beta}S_{\beta}. \tag{6}$$

Estimating the ReTF by substituting (6) into (4) gives us

$$\frac{M_1(f,t)}{M_2(f,t)} = \frac{H_{1\alpha}S_{\alpha} + H_{1\beta}S_{\beta}}{H_{2\alpha}S_{\alpha} + H_{2\beta}S_{\beta}} = \mathcal{F}(f,t,S_{\alpha},S_{\beta}), \tag{7}$$

which is an unknown function $\mathcal{F}$ that is no longer independent of the source's signal content, and therefore, no longer a useful spatial feature of either sound source. In this sense, the ReTF is not generalizable to multiple simultaneous sound sources. This problem holds even when increasing the number of receivers [2].

3. THE RELATIVE TRANSFER MATRIX

Our aim is to find a spatial property analogous to the ReTF that is generalizable to multiple simultaneous sound sources. To this end we propose formulating the ReTF as a matrix.

Consider now Q receivers indexed $q = \{1, \cdots, Q\}$ and $\mathcal{L}$ sound sources indexed $\ell = \{1, \cdots, \mathcal{L}\}$. Let us separate the receivers into two subgroups denoted by $\{A\}$ and $\{B\}$, assigned with Q_A and Q_B receivers, respectively. Figure 2 illustrates one example with $Q = 6$, $Q_A = 2$, $Q_B = 4$, and $\mathcal{L} = 2$ where $\ell = \{\alpha, \beta\}$. We express the signals received by each microphone group in matrix form as

$$\boldsymbol{M}_A(f,t) = \boldsymbol{H}_A(f)\boldsymbol{S}(f,t), \tag{8}$$
$$\boldsymbol{M}_B(f,t) = \boldsymbol{H}_B(f)\boldsymbol{S}(f,t), \tag{9}$$

where $\boldsymbol{M}_A = [M_1, \cdots, M_{Q_A}]^T$, $\boldsymbol{S} = [S_1, \cdots, S_{\mathcal{L}}]^T$, $[\cdot]^T$ is matrix transpose, and $\boldsymbol{H}_A \in \mathbb{C}^{Q_A \times \mathcal{L}}$ is a matrix with elements defined by the acoustic transfer functions.

The vector $\boldsymbol{M}_B \in \mathbb{C}^{Q_B \times 1}$ and the matrix $\boldsymbol{H}_B \in \mathbb{C}^{Q_B \times \mathcal{L}}$ are similar.

Following the mapping property of the ReTF in (3), we intend to find a matrix $\boldsymbol{\mathcal{R}}_{A,B}(f)$ that defines the spatial mapping between the receiver groups- $\{A\}$ and $\{B\}$, such that

$$\boldsymbol{M}_A(f,t) = \boldsymbol{\mathcal{R}}_{A,B}(f)\boldsymbol{M}_B(f,t). \tag{10}$$

We term $\boldsymbol{\mathcal{R}}_{A,B}(f)$ the *relative transfer matrix*. The theoretical definition of the ReTM is found by multiplying (9) by a suitable pseudo-inverse of $\boldsymbol{H}_B$ and substituting for $\boldsymbol{S}$ in (8), resulting in

$$\boldsymbol{\mathcal{R}}_{A,B}(f) = \boldsymbol{H}_A(f)\boldsymbol{H}_B^{\dagger}(f), \tag{11}$$

where $(\cdot)^{\dagger}$ denotes Moore–Penrose inverse, assuming it is valid. Just like the ReTF, again we observe that the source signals elegantly cancel off, leaving behind a spatial feature of the sources.

The ReTM (11) is seen to be a matrix defined solely by the spatial properties (the transfer functions) of the sound sources. Furthermore, we comment that the ReTM has the following three key properties analogous to the ReTF:

- The ReTM is independent of the emitted source signals,
- It is a spatial function of the sound sources, described by their acoustic transfer functions,
- As we show next, the ReTM can be estimated directly from received signals, without measuring acoustic transfer functions.

3.1. Estimating the ReTM from received signals

There are likely many ways to estimate or model the ReTM directly from received signals. In this work, we explore one method using the covariance matrices of

$$\boldsymbol{\mathcal{P}}_{AA}(f) \triangleq E\{\boldsymbol{M}_A \boldsymbol{M}_A^*\}, \text{ and } \boldsymbol{\mathcal{P}}_{BA}(f) \triangleq E\{\boldsymbol{M}_B \boldsymbol{M}_A^*\}, \quad (12)$$

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

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{AA}(f) &\cong \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \boldsymbol{M}_A(f,t)\boldsymbol{M}_A^*(f,t) \\ \boldsymbol{\mathcal{P}}_{BA}(f) &\cong \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \boldsymbol{M}_B(f,t)\boldsymbol{M}_A^*(f,t). \end{aligned} \quad (13)$$

Using (8) and (9) in (12), we write

$$\boldsymbol{\mathcal{P}}_{AA}(f) = \boldsymbol{H}_A \boldsymbol{\mathcal{P}}_S \boldsymbol{H}_A^* \quad (14)$$

$$\boldsymbol{\mathcal{P}}_{BA}(f) = \boldsymbol{H}_B \boldsymbol{\mathcal{P}}_S \boldsymbol{H}_A^* \quad (15)$$

where $\boldsymbol{\mathcal{P}}_S \triangleq E\{\boldsymbol{S}\boldsymbol{S}^*\}$ is the expectation of the source signals. By multiplying (15) by $\boldsymbol{H}_B^\dagger$ and substituting into (14), we obtain

$$\boldsymbol{\mathcal{P}}_{AA}(f) = \boldsymbol{H}_A \boldsymbol{H}_B^\dagger \boldsymbol{\mathcal{P}}_{BA}(f) = \boldsymbol{\mathcal{R}}_{A,B}(f)\, \boldsymbol{\mathcal{P}}_{BA}(f). \quad (16)$$

The ReTM is estimated by applying the pseudo-inverse of $\boldsymbol{\mathcal{P}}_{BA}(f)$ to (16) as

$$\boldsymbol{\mathcal{R}}_{A,B}(f) = \boldsymbol{\mathcal{P}}_{AA}(f) \boldsymbol{\mathcal{P}}_{BA}^\dagger(f). \quad (17)$$

In Practice, microphones will have additive thermal noise, thus the estimation of ReTM in (17) is only an approximation. This is also the case for estimating the ReTF using power spectral densities. We investigate the conditioning of covariance matrix $\boldsymbol{\mathcal{P}}_{BA}$ and thereby the practicality of the ReTM with experimental recordings next.

4. PRELIMINARY ANALYSIS

In this section, we provide a preliminary validation of the ReTM under the following three analysis scenarios:

1. *ISM*: Numerical simulations using image source method;
2. *IR*: Experimental recordings from measured impulse responses;
3. *Live*: Experimental recordings from live loudspeakers.

There are many independent variables to validate including the number of sources and microphones. Here, we only consider the case of two sources and 6 microphones as shown in Fig. 2 to gain an initial understanding of the ReTM. Figure 3 provides an approximate illustration of the setup in each scenario. We inspect the condition number of covariance matrix $\boldsymbol{\mathcal{P}}_{BA}$, as well as the average magnitude spectrum error of the first (q=1) microphone in group A given by

$$\mathcal{E}_1(f) = \underset{t}{\text{mean}}\, 10\log_{10} |\hat{M}_1(f,t) - M_1(f,t)|^2 / |M_1(f,t)|^2,$$

where $\hat{M}_1(f,t)$ is the estimated signal at q=1 microphone using the ReTM in $\mathcal{R}_{A,B}\boldsymbol{M}_B$. In a sense, $\hat{M}_1$ can be thought of as a remote microphone signal that is estimated from the ReTM and the signals received in group B. If the remote estimate of $\hat{M}_1$ matches the original recorded signals of M_1 then the ReTM can be considered to be mapping ReTFs of multiple sources as intended. Furthermore, if this remote estimate remains accurate as the source signals change, then the ReTM is independent of the emitted signals and is a spatial property of the sound sources.

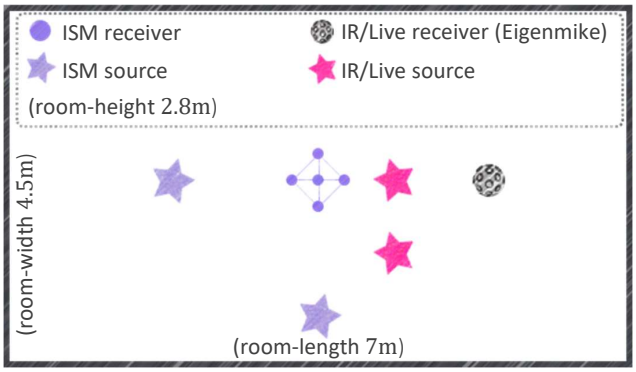

Fig. 3. Approximate setup of the receiver and source positions in the numerical simulation and experiment recordings.

4.1. Numerical simulations

In brief, we model the acoustic transfer function between a source at (x_ℓ, y_ℓ, z_ℓ) and receiver at (x_q, y_q, z_q) with respect to the front-left-bottom corner of a shoe-box room with [25]

$$H_{q\ell}(f) = \sum_{\boldsymbol{w}=0}^{1} \sum_{\boldsymbol{d}=-D}^{D} \delta_{x1}^{|d_1-w_1|} \delta_{x2}^{|d_1|} \delta_{y1}^{|d_2-w_2|} \delta_{y2}^{|d_2|} \delta_{z1}^{|d_3-w_3|} \delta_{z2}^{|d_3|} \times e^{i\frac{2\pi f}{c}|\vec{x}_w + \vec{x}_d|} / 4\pi |\vec{x}_w + \vec{x}_d|, \quad (18)$$

where $D(=8)$ is image depth, c is speed of sound, $\sum_{\boldsymbol{w}/\boldsymbol{d}}$ denote triple summations over the index terms $\boldsymbol{w}=(w_1,w_2,w_3)$ and $\boldsymbol{d}=(d_1,d_2,d_3)$, $\delta(=0.9)$ are reflection coefficients of the close ($\{1\}$) and far ($\{2\}$) walls along each Cartesian axis, $\vec{x}_w=(x_q-x_\ell+2w_1x_\ell,\ y_q-y_\ell+2w_2y_\ell,\ z_q-z_\ell+2w_3z_\ell)$, and $\vec{x}_d=(2d_1 \times \textit{room-length},\ 2d_2 \times \textit{room-width},\ 2d_3 \times \textit{room-height})$.

The six receivers are configured as an octahedron ± 0.2 m along each axis, the two on the z-axis comprise group A, the remaining are group B. The signals are processed directly in the short-time Fourier domain for a 512 window size, 48 kHz sampling, and 10 second duration. The sources both emit white noise.

4.2. Experimental recordings

Both the *IR* and *live* recordings shared the same setup. We used an em32 Eigenmike [26] for the recordings with 48 kHz sampling. We assigned the two receivers on the back (channels $\{18, 20\}$) to group A, and the 4 receivers on the front (channels $\{1\text{-}4\}$) to group B. We note here that the Eigenmike is a spatially small and symmetric device. It is expected that a more spatially diverse microphone configuration would perform better in a ReTM application. The sources were two loudspeakers playing vacuum cleaner noise and music, placed ~1 m in front of the microphones. We note that these stimulus signals were compressed by a 16 kHz low-pass, which limits our analysis. The room was a large office with minor acoustic treatment on the walls, a $T_{60} \approx 200$ ms reverberation time, and noteworthy background air conditioning noise (~45 dBA noise floor). The 10 second time domain signals were transformed into the short-time Fourier domain with a 2^{14} window size that is longer than the reverberation time to satisfy the multiplicative transfer function [27].

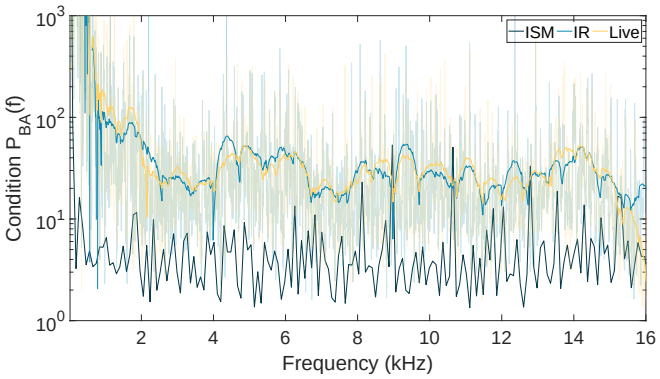

Fig. 4. Condition number of the $\boldsymbol{\mathcal{P}}_{BA}(f)$ covariance matrix.

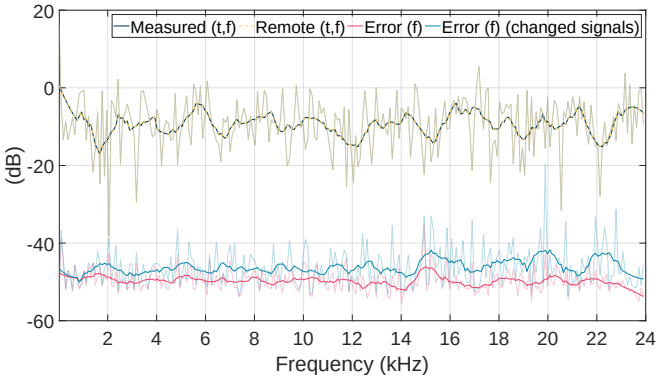

Fig. 5. Magnitude spectrum and error averaged over time for the *ISM* numerical simulation.

4.3. Results

Figure 4 provides the condition number of the covariance matrix $\boldsymbol{\mathcal{P}}_{BA}(f)$ estimated by (13) in each scenario. A low condition number indicates that the ReTM can be estimated via (17) while being robust to erroneous noise. We observe that the *ISM* simulation has good condition of mostly $<$10 throughout the wide frequency band. Whereas, the *IR* and *live* recordings are conditioned similar to each other, but worse than the *ISM*, with values between 10-100 above ~2 kHz. At low frequencies, the ill-conditioning may be a result of the Eigenmike's closely spaced receivers.

Figure 5 gives the *ISM* magnitude spectrum of the recorded M_1 and estimated remote $\hat{M}_1$ signals (at a single time-frame) and their error averaged over all time frames. Immediately, we observe an almost perfect match between the measured and remote signals that are estimated through via $\boldsymbol{M}_B$ and the ReTM using (10). Below -40 dB error occurs throughout the full (smoothed) 20-24k Hz band. This suggests that the ReTM estimated by (17) models the spatial mapping between microphone groups A and B correctly, just like the ReTF. To evaluate the ReTM's independence of source signal we repeat the remote signal estimation (10) with the same ReTM, where now the second source has changed from white noise to rainfall sound. That is, the ReTM is calculated by (17) when the source is white noise, and then used to estimate the remote signals when the source changes signals. In this case, the results in Fig. 5 denoted by "changed signals" are good. The magnitude spectrum error of the remote signals remains low, indicating that the ReTM is correctly mapping each source's spatial property while being independent of their signals.

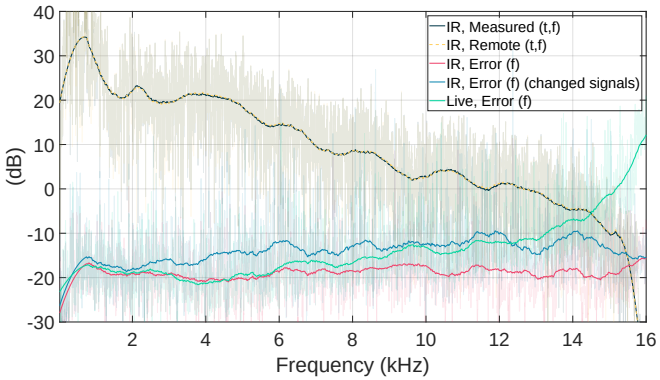

Fig. 6. Magnitude spectrum and error averaged over time for the *IR* and *live* experiment recordings.

Figure 6 shows the magnitude spectrum and error again, this time for the *IR* and *live* experiment recording scenarios. Once more, we observe a close match between the measured and remote signals. As to be expected, the real-world measurements have worse error than the ideal simulations. However, the *IR* error is still below -10 dB for the full (smoothed) 20-16k Hz band. Furthermore, we see a low error below 2 kHz despite the ill-conditioning we observed in Fig. 4. The *live* recording error is also similarly good, remaining below -10 dB for most frequencies, but becoming worse above 12 khz. This may be due to the source's energy being low compared to the non-stationary room noise. Finally, we change the source signals in the *IR* scenario after the ReTM is estimated. The result (denoted by "changed signals") is for the first source changing to music and the second changing to rainfall sound. Similar to *ISM*, the error slightly worsens from changing source signals. However, it continues to remain below -10 dB for all frequencies, supporting that the ReTM is independent of source signals.

5. CONCLUSION

Thus far, we have only provided a preliminary proof-of-concept examination of the ReTM. We leave a more thorough investigation to its implementation in a conventional spatial acoustic problem as future work. We note here that we have not provided any rule-of-thumb for the required number of receivers and their grouping for a given number of sources. Informally, we found that for three simultaneous sound sources, we required three microphones in group A. While increasing the number of microphones in group B provided unremarkable benefits.

Nonetheless, we have proposed a generalization to the relative transfer function in the form of the *relative transfer matrix*. By separating receivers into two multichannel groups, we are able to model their coupled response due to multiple sound sources. We showed that the ReTM can be estimated from observed signals through a covariance based method. Furthermore, we demonstrated that the ReTM is a spatial property that is independent of the source signals in experimental measurements. In this regard, the ReTM behaves functionally like the ReTF while being generalized for multiple sources. There remains substantial promise in the capabilities of the ReTM model. Ideally, existing uses of the ReTF can be naturally extended to multi-source scenarios. Multiple source localization, tracking, and separation are one possible example.

6. REFERENCES

[1] Maja Taseska and Emanuel A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 404–408.

[2] 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.

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

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

[5] 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.

[6] G. Reuven, S. Gannot, and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 4, pp. 711–727, 2008.

[7] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 11–15.

[8] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 320–324.

[9] 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.

[10] 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.

[11] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[12] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 425–437, 1997.

[13] N. Gößling, W. Middelberg, and S. Doclo, "Rtf-steered binaural mvdr beamforming incorporating multiple external microphones," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 373–377.

[14] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[15] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[16] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, "Advances in network and acoustic echo cancellation," *Springer*, 2001.

[17] A. Sofer, T. Kounovskỳ, J. Čmejla, Z. Koldovskỳ, and S. Gannot, "Robust relative transfer function identification on manifolds for speech enhancement," in *Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2021, pp. 401–405.

[18] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6500–6504.

[19] B. Yang, H. Liu, and X. Li, "Learning deep direct-path relative transfer function for binaural sound source localization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3491–3503, 2021.

[20] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2171–2186, 2016.

[21] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Sound source localization using relative harmonic coefficients in modal domain," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 348–352.

[22] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using relative harmonic coefficients," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 571–575.

[23] Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Decoupled multiple speaker direction-of-arrival estimator under reverberant environments," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 30, pp. 3120–3133, 2022.

[24] O. Shalvi and E. Weinstein, "System identification using non-stationary signals," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2055–2063, 1996.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[26] MH Acoustics, "Em32 eigenmike microphone array release notes (v17. 0)," *25 Summit Ave, Summit, NJ 07901, USA*, 2013.

[27] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, 2007.