On the use of the Relative Transfer Function for Source Separation using Two-channel Recordings

Alice P. Bates, Daniel Grixti-Cheng, Prasanga Samarasinghe and Thushara Abhayapala
Research School of Electrical, Energy and Materials Engineering, Australian National University
E-mail: prasanga.samarasinghe@anu.edu.au

***Abstract*—This paper investigates the use of the relative transfer function (ReTF) for source separation. ReTFs have been used to localize sound sources but have not been thoroughly investigated for the application of source separation especially when one of the sources is not speech. We propose two source separation algorithms using ReTFs. One of them is deterministic and enables the separation of two sources when one or both of their ReTFs are known. The other algorithm uses masking in the time-frequency domain and can be used for separating two or more sources. We also explore the limitations and assumptions of the ReTF and the proposed source separation algorithms.**

***Index Terms*—Source separation, relative transfer function, time-frequency masking**

I. Introduction

The relative transfer function (ReTF) between microphones represents the coupling between these microphones in response to a source. The ReTF gives a unique signature of the source position, as well as the position of the microphones and the environmental characteristics, such as, room dimensions and reverberation time [1]–[4]. Hence the ReTF contains very useful information and is commonly used for determining source location [5]–[7]. These source localization methods often require many microphones, however in practice, most devices are equipped only with a small number of microphones to save hardware cost and computational complexity.

We investigate the ReTF for the application of source separation on two-channel recordings. An example of a practical application of this is in teleconferencing where we are interested in extracting speech from background noise such as the air-conditioning in the office or the traffic going by outside the window. Little has been done in this area; a few methods exist [8]–[10] which use features of two-channel recordings in the short-time Fourier transform (STFT) domain which are related to the ReTF. These methods separate a mixture of speech signals using the Watson mixture model for clustering across frequency bins and an additional step for assigning clusters to the different sources present in the recording. These algorithms only work on separating mixtures of speech signals.

The vast majority of source separation algorithms [4], [11]–[13], [13] use supervised machine learning techniques [14]–[17]. These methods require large amounts of labeled data which is not practical in many source separation problems. We consider the case where there is only a small amount of data available and little information known about the recordings; we do not know environment characteristics, such as, dimensions, reverberation time and location of sources or microphones.

II. Relative transfer function features

A. Problem Fomulation

Consider recorded signals from two microphones, $y_A(t)$ and $y_B(t)$ placed in a room with N sound sources. Let signals emitted by sound sources be $s_i(t)$, $i = \ldots, N$ and h_{iA} and h_{iB} be the room impulse response (RIR) from the ith source to the two microphones. We can write the received signal at the microphone A or B as [3], [17]

$$y_{i\{A,B\}}(t) = \sum_{i=1}^{N} h_{\{A,B\}}(t) * s_i(t). \tag{1}$$

where $\{\cdot\}$ denotes the microphone of interest, A or B. We take STFT of (1), provided the time-window length used in STFT is long relative to the length of the RIR, the multiplicative transfer function model (MTF) [18] applies with the ReTF only being a function of frequency, not time. Thus,

$$Y_{\{A,B\}}(p,k) = \sum_{i=1}^{N} H_{i\{A,B\}}(k) S_i(p,k), \tag{2}$$

where H_{iA} and H_{iB} are the acoustic (room) transfer function (RTF) from the ith source to the two microphones, and p and k denote the time and frequency indexes, respectively. Source separation is the process of taking a recording with a mixture of sources (1) or (2) and recovering a single source $h_{iA}(t) * s_i(t)$ or $H_{iA}(k)S_i(p,k)$. We do not consider dereverberation in this work.

B. Single source relative transfer function

For a single source, the ReTF $R(k)$ is defined as the ratio between the RTF of microphone A, and the RTF of microphone B, $R_1(k) \triangleq H_{1B}(k)/H_{1A}(k)$ with respect to a source. When there is only one source is present

$$R_i(k) = \frac{Y_B(p,k)}{Y_A(p,k)} = \frac{H_{iB}(k)S_i(p,k)}{H_{iA}(k)S_i(p,k)} \tag{3}$$

There are a plethora of algorithms to estimate $R(k)$ with microphone noise [7]. We use $R(k) \approx \Phi_{y_A y_B}(k)/\Phi_{y_A y_A}(k)$, where $\Phi_{y_A y_B}(k)$ and $\Phi_{y_A y_A}(k)$ are the cross-power and auto-power spectral density of the audio recordings respectively.

Note that the ReTF is independent of the source signal but provides a uniques signature of the source spatial

characteristics for a given room and microphone locations.

C. Multi-source relative transfer function

We extend the ReTF definition in (3) to multiple sources. Using the MTF model, the ReTF for N sources is given by

$$R(p,k) = \frac{Y_B(p,k)}{Y_A(p,k)} = \frac{\sum_{i=1}^{N} H_{iB}(k)S_i(p,k)}{\sum_{i=1}^{N} H_{iA}(k)S_i(p,k)}. \quad (4)$$

If the time-window of the STFT is chosen small enough, then the W-disjoint [19] assumption holds. The W-disjoint assumption states that only one source signal is present in each time-frequency bin of the STFT. This means that the ReTF for a time bin α and a frequency bin β equals

$$R(p_\alpha, k_\beta) = \frac{H_{iB}(k_\beta)S_i(p_\alpha,k_\beta)}{H_{iA}(k_\beta)S_i(p_\alpha,k_\beta)} = \frac{H_{iB}(k_\beta)}{H_{iA}(k_\beta)} = R_i(k_\beta). \quad (5)$$

That is, each time-frequency bin of an ReTF calculated for multiple sources in the time-frequency domain, is equal to a frequency bin of the ReTF of a single source.

If the time-window length of the STFT is small compared with the length of the RIR, the convolutive transfer model (CTF) [2], [4], rather than the MTF, is valid and the ReTF is a function of both frequency and time with

$$R(p,k) = \frac{Y_B(p,k)}{Y_A(p,k)} = \frac{\sum_{p'=0}^{Q}\sum_{i=1}^{N} H_{iB}(p',k)S_i(p-p',k)}{\sum_{p'=0}^{Q}\sum_{i=1}^{N} H_{iA}(p',k)S_i(p-p',k)}, \quad (6)$$

where Q is the length of the RIR divided by the length of the time-frame used in the STFT. Using the CTF and the W-disjoint assumption, the ReTF for a time bin α and a frequency bin β equals

$$\begin{aligned} R(p_\alpha,k_\beta) &= \frac{\sum_{p'=0}^{Q} H_{iB}(p',k_\beta)S_i(p_\alpha-p',k_\beta)}{\sum_{p'=0}^{Q} H_{iA}(p',k_\beta)S_i(p_\alpha-p',k_\beta)} \\ &= R_i(p_\alpha,k_\beta). \end{aligned} \quad (7)$$

That is, each time-frequency bin of an ReTF calculated for multiple sources in the time-frequency domain, is equal to a time-frequency bin of the ReTF of a single source in the time-frequency domain.

III. Source separation algorithms

In this section, we present two source separation algorithms we have developed, both of which use the ReTF. The first uses the single source ReTF (II-B) and is a deterministic algorithm for separating two sources. The second uses the ReTFs for single (II-B) and multiple (II-C) sources to estimate binary masks for separating two or more source signals. The single source ReTFs have to be precomputed from measurements done when only one source is active. For example, in a conference room, recordings of the air-conditioner can be made when the room is idle.

A. Deterministic two source separation

For $N = 2$ sources, the audio signal received at microphones A and B can be written as (2)

$$Y_A(p,k) = H_{1A}(k)S_1(p,k) + H_{2A}(k)S_2(p,k) \quad (8)$$

$$Y_B(p,k) = H_{1B}(k)S_1(p,k) + H_{2B}(k)S_2(p,k). \quad (9)$$

Using (3), (9) can also be written as

$$Y_B(p,k) = H_{1A}(k)R_1(k)S_1(p,k) + H_{2A}(k)R_2(k)S_2(p,k). \quad (10)$$

If $R_2(k)$ is known, using (10) and (8) we can compute

$$Y_B(p,k) - R_2(k)Y_A(p,k) = \big(R_1(k) - R_2(k)\big)H_{1A}(k)S_1(p,k) \quad (11)$$

In the right hand side of (11), source 2, $S_2(p,k)$, is removed and what remains is a filtered version of source 1. If both ReTFs are known, this filtering can be removed to recover the audio signal received at microphone A if only source 1 is present, $Y_A(p,k) = H_{1A}(k)S_1(p,k)$. ReTFs of one or both sources are estimated using (3) during periods of the recordings when only one source is active.

B. Time-frequency masking multi-source separation

We wish to form binary masks $\mathcal{M}_i(p,k)$ in the time-frequency domain to recover an estimate of each source $\hat{S}_i(p,k)$ with,

$$\hat{S}_i(p,k) \triangleq \mathcal{M}_i(p,k)Y_A(p,k), \quad \forall i = 1,...,N. \quad (12)$$

The masks can be recovered by working out which source at each time-frequency bin (p_α, k_β) of the multi-source ReTF comes from which individual source's ReTF. However, our efforts of clustering each frequency bin for all time-frames is not successful and we find that the multi-source ReTF values for a frequency bin does not form separable clusters in practice. Thus, we conclude that the MTF model (5) does not hold perfectly in practice.

The CTF model (7) is source and time dependent. If we have the single source ReTF in the time-frequency for each source, then we can separate sources by working out for each time-frequency bin of the multi-source ReTF which of the time-frequency bins of the individual sources it is closest too. Unfortunately, for real recordings we cannot compute the individual sources' ReTFs in the time-frequency domain. Thus, he CTF model (7) is not practical as each source has to be the same on its own as in the mixture.

1) Bin-wise method: It is possible to find the ReTF of each individual source in the frequency domain $R_i(k)$, provided there are periods in the recording where each source is present on its own. We can then compare each time-frequency bin of the multi-source ReTF to the same frequency bin of the single source ReTFs and assign the time-frequency bins to whichever source had the closer ReTF value for that frequency. That is, for each time bin α and frequency bin β the source I which will have the bin assigned to its mask $\mathcal{M}_I(p,k)$ is

$$I = \arg\min_i \Big\{ ||R(p_\alpha,k_\beta) - R_i(k_\beta)|| \Big\}, \quad (13)$$

where $||.||$ denotes the Euclidean norm.

2) Multi-bin method: A limitation with th above bin-wise assignment is that the ReTF is a unique function of frequency but not necessarily unique for a single frequency bin. It is therefore difficult to compare ReTFs in a bin-wise fashion. Since multiple frequency bins in the time-frequency domain tend to be active synchronously [10], (known as the common amplitude modulation property), we can look at multiple neighbouring frequency bins together. The number of bins within a bin radius R that are closer to source I is given by

$$C = \sum_{b=\beta-R}^{\beta+R} ||R(p_\alpha, k_\beta) - R_X(k_\beta)|| < \min_i ||R(p_\alpha, k_\beta) - R_{i\neq X}(k_\beta)||. \quad (14)$$

The time-frequency bin is assigned to the mask for source I if the ratio $C/(2R+1)$, is greater than the specified dominance ratio D with,

$$\mathcal{M}_I(p_\alpha, k_\beta) = \begin{cases} 1 & \text{if } C/(2R+1) \geq D \\ 0 & \text{if } C/(2R+1) < D. \end{cases} \quad (15)$$

IV. Performance analysis

In order to analyze the performance of the source separation algorithms presented in Section III, we conduct the following source separation experiments for the measurement setup shown in Fig. 1 and reverberation times in the range $[100, 450]$ms. In Fig. 1 a female news presenter is contaminated by a much louder undesired sound source. We use traffic noise (TN) as an example of a broadband source and air-conditioning (AC) as an example of a narrow band source. The undesired sound sources have 160 times the energy of the desired female news presenter source. The sampling frequency of all sources and recordings is 16kHz, and 40dB white Gaussian microphone noise is added to all the recordings.

A common approach to evaluating the quality of an estimated source signal is to compute the overall perceptual score (OPS) which is an energy ratio between the reference, i.e., the clean target signal, and that of the estimation. We calculate the OPS using the PEASS Toolbox [20].

A. Deterministic two source separation method

Here we evaluate the deterministic two source separation method described in Section III-A for when just the ReTF of the undesired source is known and when the ReTF of both sources are known. The OPS for both of these cases as a function of reverberation time and for two distances of the desired source from the microphones is shown in Fig. 2. In Fig. 2 the OPS is close to 100dB when both ReTFs are known indicating near perfect reconstruction as expected. The OPS

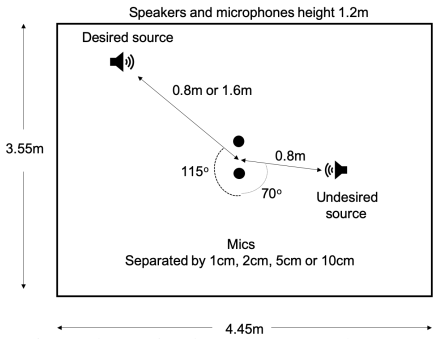
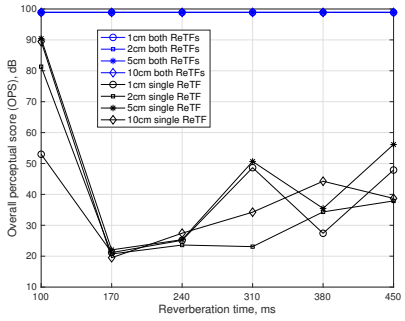
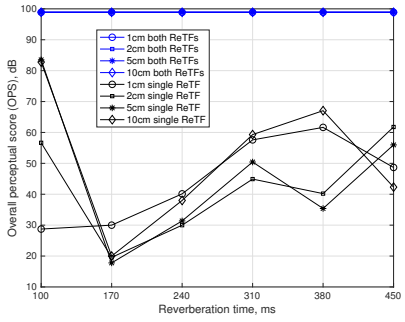


Fig. 1: Configuration of microphones and sources used for experimental evaluations.

is much lower when only one ReTF is known due to the distortion introduced in (11). For the shortest reverberation time of 100ms, larger microphone separation distances and shorter distance of the desired source from microphones leads to better OPS but there is little difference in OPS at longer reverberation times. For the deterministic two source separation method, the separated signal, given by (11), is independent of the undesired source signal, only on the ReTFs of the two sources, hence the same OPS is achieved for either TN or AC as the undesired sound source.



(a)

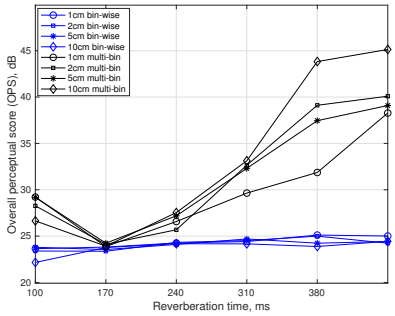


(b)

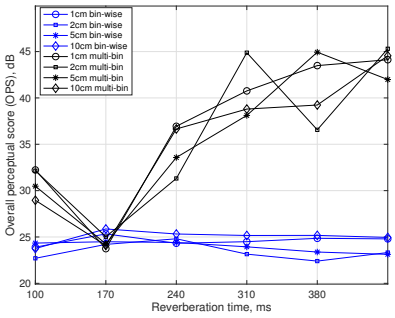
Fig. 2: Overall perceptual score (OPS) for deterministic two source separation method for female news presenter (FNP) being recovered from recording with undesired sound source for FNP distance 0.8m (a) and 1.6m (b) from microphones.

B. Time-frequency masking multi-source method

Here we include a preliminary evaluation of the multi-source separation method described in Section III-B. When evaluating the multi-bin approach in Section III-B2. For the experimental setup in Fig. 1, a STFT window size of 2048 samples (128ms) and frame shift of 512 (32ms), a bin radius $C = 1$ and dominance ratio $D = 1$ achieved the best results. The OPS calculated using the bin-wise and the multi-bin approach (with $C = 1$ and $D = 1$) are shown in Fig. 3, and as a function of reverberation time and for two distances of the desired source from the microphones. The multi-bin approach performs better than the bin-wise approach for all reverberation times and speaker distances expect for a single reverberation time (170ms). The same OPS was achieved for either TN or AC as the undesired sound source.



(a)



(b)

Fig. 3: Overall perceptual score (OPS) for the time-frequency masking multi-source separation method for female news presenter (FNP) being recovered from recording with undesired sound source for FNP distance 0.8m (a) and 1.6m (b) from microphones.

V. Existing challenges

For both the deterministic two source separation and the time-frequency masking multi-source separation method, a limitation is having to compute ReTFs of individual sources. In order to do this, there needs to be parts of the recording where each source is present on its own. For the deterministic two source separation method, although only one source's ReTF is required, the method is limited to two sources.

For the multi-source separation method, it is not possible to get the STFT window length so that both the W-disjoint assumption and the MTF assumption both hold exactly, so there is always a trade-off between the two assumptions. If the recording only contains signals that are sparse in time, such as speech signals, it is possible to have a longer window and still maintain the W-disjoint assumption so that both assumptions approximately hold.

The multi-source separation method can in theory separate any number of sources provided their ReTFs can be learned, and the W-disjoint and the MTF assumptions approximately hold. As the number of sources increases, the W-disjoint assumption increasingly breaks down [19]. We are investigating comparing multiple bins in time as well as frequency to increase the robustness of this method. In addition, the best bin radius and dominance ratio to use is unclear and might depend on the individual recording environment, this needs to be investigated further. We found that larger bin radii did not work well for source separation as few time-frequency bins were set to 1 in the binary mask.

VI. Conclusion

In this paper we investigated the use of the relative transfer function (ReTF) for source separation. We proposed two source separation algorithms: one which is deterministic and enables the separation of two sources when one or both of their ReTFs is known and the other algorithm which uses masking of the ReTF in the time-frequency domain to enable multi-source separation. We also explored the assumptions of the ReTF and analyze the performance of the proposed source separation algorithms. We conclude, though there are limitations that need to be taken into account, that the ReTF is a promising feature for source separation.

References

[1] 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, Aug. 2001.

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

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

[4] L. Li, H. Kameoka, and S. Makino, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 546–550.

[5] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," New Platz, NY, Oct. 2013, pp. 1–4.

[6] ——, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 8, pp. 1393–1407, Aug. 2016.

[7] ——, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 7, pp. 1477–1491, Jul. 2017.

[8] I. Jafari, R. Togneri, and S. E. Nordholm, "On the use of the Watson mixture model for clustering-based under-determined blind source separation," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, Sep. 2014, pp. 988–992.

[9] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 3238–3242.

[10] ——, "Permutation-free clustering of relative transfer function features for blind source separation," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp. 409–413.

[11] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "PSD estimation and source separation in a noisy reverberant environment using a spherical microphone array," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1594–1607, 2018.

[12] S. Makino, *Audio Source Separation*. Springer, 2018.

[13] N. Chong, S. Nordholm, B. T. Vo, and I. Murray, "Tracking and separation of multiple moving speech sources via cardinality balanced multi-target multi bernoulli (cbmember) filter and time frequency masking," in *2016 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2016, pp. 88–93.

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

[15] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[16] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, C.-A. Deledalle, and W. Li, "Machine learning in acoustics: a review," *arXiv e-prints*, p. arXiv:1905.04418, May 2019.

[17] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018.

[18] 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, May 2007.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[20] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.