Relative Transfer Function Estimation From Speech Keywords

Ryan M. Corey* and Andrew C. Singer

University of Illinois at Urbana-Champaign

Abstract. Far-field speech capture systems rely on microphone arrays to spatially filter sound, attenuating unwanted interference and noise and enhancing a speech signal of interest. To design effective spatial filters, we must first estimate the acoustic transfer functions between the source and the microphones. It is difficult to estimate these transfer functions if the source signals are unknown. However, in systems that are activated by a particular speech phrase, we can use that phrase as a pilot signal to estimate the relative transfer functions. Here, we propose a method to estimate relative transfer functions from known speech phrases in the presence of background noise and interference using template matching and time-frequency masking. We find that the proposed method can outperform conventional estimation techniques, but its performance depends on the characteristics of the speech phrase.

Keywords: Relative transfer function, multichannel source separation, keyword spotting, microphone array

1 Introduction

In many audio processing applications, such as voice assistants and augmented listening devices, we wish to isolate a single speech signal of interest from background noise and interference. These systems can use microphone arrays to spatially filter audio signals, emphasizing sounds from a target direction and attenuating signals from other directions [1]. Multichannel processing has been shown to improve the performance of speech recognition systems in noisy environments [2]. Arrays can also be used in hearing aids and other listening devices to enhance human hearing [3]. In order to filter out interference, a system must determine which signals are coming from which source. We can differentiate sources using their relative transfer functions (RTF), which describe differences in sound propagation between sources and microphones and are generally different for sources in different locations [4]. In environments with significant reverberation, particularly when devices are placed next to walls or other reflecting surfaces, the RTFs are difficult to predict geometrically and must be estimated from observed data.

* Parts of this research were completed at Arm Research. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245.

Because the target speech and unwanted interference signals are generally unknown, RTF estimation is a difficult problem. If the desired signal is stronger than the background noise or if the noise statistics can be reliably estimated, then the RTFs can be estimated using subspace techniques [5, 6]. The RTFs can also be estimated using a variety of blind source separation techniques that rely on assumptions about the properties of the signals [7, 8]. It would be more reliable to estimate RTFs using a known pilot signal like those used in communication systems [9], but such signals are typically unavailable. In some applications, however, we do have partial knowledge of the content of the speech. We can therefore use the speech itself as a pilot signal to estimate the RTFs.

In this work, we consider audio capture systems that are activated by a certain speech phrase, known as a keyword. Such keywords are often used to remotely activate voice assistants on mobile phones and other electronic devices. These systems use low-power keyword spotting algorithms to continuously monitor for the speech phrase, then activate the full recognition system once it is detected [10]. Because the content of this speech phrase is known in advance, we can use the keyword to better estimate the RTFs of the speaker. Here, we propose an RTF estimation system that matches a multichannel recording to a prerecorded template of the keyword, uses that template to isolate the keyword in each channel, and estimates the RTFs from those isolated recordings. To demonstrate the source separation utility of the keyword alone, we do not apply any other blind source separation techniques and we use no information about the array geometry. A key question in this study is the impact of the choice of keyword on the performance of the system: how do the length and spectral content of the keyword affect the accuracy of the RTF estimate? We will demonstrate the performance of the system and address this question using a crowdsourced database of speech commands and a microphone array similar to those used in commercial voice-assistant-enabled speakers.

2 Far-Field Audio Capture

A far-field audio capture system is shown in Figure 1. Sound is captured by an array of M microphones, which we assume to behave linearly but which may have arbitrary locations and frequency responses. The system continuously records from all M microphones while it waits for the keyword. The signals are processed as follows:

1. A keyword spotting algorithm, which we assume to work perfectly, activates the system upon detecting the keyword.
2. Once the system is activated, the keyword is used to estimate the relative transfer functions of the source.
3. The RTFs are used to design a source separation filter that isolates the speech following the keyword and suppresses interference and noise.
4. The separated speech is then reproduced, stored, transmitted, or processed further, depending on the application.

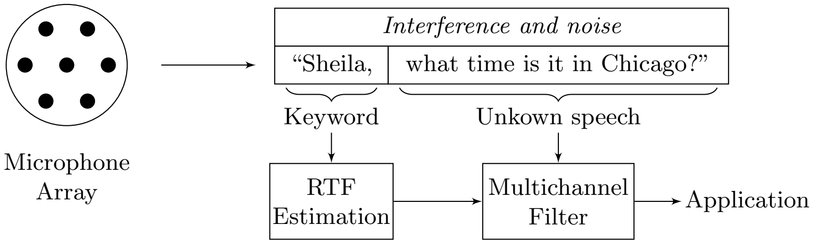

Fig. 1. A far-field audio capture system uses a known speech keyword to design a multichannel filter and separate the unknown speech.

2.1 Microphone Array System

Let $s(n,k)$ be the short-time Fourier transform (STFT) of the signal of interest at the first microphone, where n is the frame index and k is the frequency index. Let $\mathbf{x}(n,k)$ be the M-dimensional STFT vector of mixture signals received by the M microphones. Under the multiplicative transfer function model [8], the mixture is given by

$$\begin{aligned}\mathbf{x}(n,k) &= \mathbf{a}(k)s(n,k) + \mathbf{z}(n,k) &(1)\\ &= \mathbf{c}(n,k) + \mathbf{z}(n,k), &(2)\end{aligned}$$

where $\mathbf{z}(n,k)$ is the M-dimensional STFT vector of unwanted interference and noise signals received by the microphones, $\mathbf{a}(k)$ is the vector of RTFs, and $\mathbf{c}(n,k) = \mathbf{a}(k)s(n,k)$ is the noise-free vector of source images. Because $s(n,k)$ is defined with respect to the first microphone, $a_1(k) = 1$ for all k. The RTFs depend on the relative positions of the source and microphones, the reverberation characteristics of the space, and the frequency responses and directionalities of the microphones, which may be unknown.

2.2 Source Separation

To isolate the signal of interest, $s(n,k)$, from the mixtures $\mathbf{x}(n,k)$, we use an M-channel spatial filter $\mathbf{w}(k)$, sometimes known as a filter-and-sum beamformer:

$$\hat{s}(n,k) = \mathbf{w}^H(k)\mathbf{x}(n,k). \tag{3}$$

There are many ways to select the coefficients. Here, we restrict our attention to the minimum power distortionless response (MPDR) coefficients [11],

$$\mathbf{w}(k) = \frac{\Sigma_x^{-1}(k)\mathbf{a}(k)}{\mathbf{a}^H(k)\Sigma_x^{-1}(k)\mathbf{a}(k)}, \tag{4}$$

where $\Sigma_x(k) = \mathbb{E}\left[\mathbf{x}(n,k)\mathbf{x}^H(n,k)\right]$ is the covariance matrix of the mixture. The MPDR filter minimizes the expected power of $\mathbf{w}^H(k)\mathbf{x}(n,k)$ while ensuring that $\mathbf{w}^H(k)\mathbf{a}(k)s(n,k) = s(n,k)$. To compute the coefficients, we must first estimate

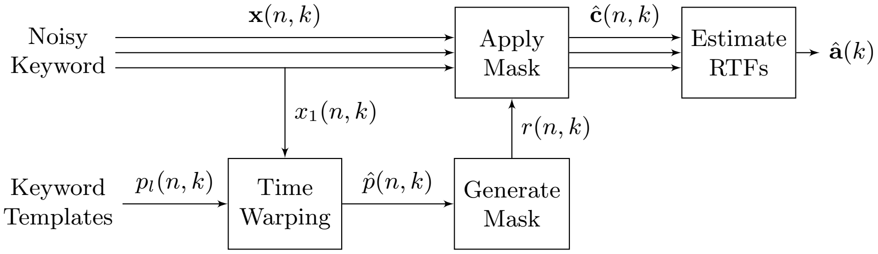

Fig. 2. The relative transfer functions are estimated from the noisy recording using a time-warped template and a time-frequency mask.

both $\Sigma_x(k)$ and $\mathbf{a}(k)$. In our experiments, the mixture covariance matrix is estimated from the recording itself,

$$\hat{\Sigma}_x(k) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}(n,k)\mathbf{x}^H(n,k). \tag{5}$$

The MPDR filter is known to be sensitive to errors in the estimate of $\mathbf{a}(k)$ [11]. While this is a disadvantage in practice, it is helpful in illustrating the RTF estimation performance of the system.

3 Relative Transfer Function Estimation

If the source and microphone positions or the room acoustics are unknown, then the RTFs must be estimated blindly from the noisy mixture data. Fortunately, in keyword-activated systems, the keyword itself can act as a pilot signal to measure the acoustic channel. Of course, the keyword signal as uttered by the speaker is not known exactly; it must itself be estimated from the noisy mixture.

The proposed method, shown in Figure 2, combines classic template matching algorithms and modern single-channel source separation methods:

1. Use dynamic time warping to match the recorded keyword to a template keyword from a database.
2. Use the warped template to generate a time-frequency mask consistent with the recorded keyword.
3. Apply the mask to each of the M channels of the mixture to isolate the recorded keyword from interference and noise.
4. Estimate the RTFs from the spatial correlation of the masked data.

To better analyze the performance of keyword-based RTF estimation and to compare different keywords, we do not apply any other blind source separation techniques and we do not use information from the speech following the keyword.

3.1 Template Matching

Template matching is a classic small-vocabulary speech recognition technique [12]. The recorded keyword signal $x_1(n,k)$ is matched to one of L templates $p_l(n,k)$ from a database. Since the sounds within a keyword can be uttered at different speeds, the templates are warped to match the time scale of the recording. Mathematically, we find the best-fitting template and the corresponding time mapping by solving the minimization problem

$$\hat{l}, \hat{t}(n) = \arg \min_{l,t(n)} \sum_n \text{Cost} \left(x_1(n,1),\ldots, x_1(n,K); p_l(t(n),1),\ldots, p_l(t(n),K) \right), \quad (6)$$

where $t(n)$ is nondecreasing. In our experiments, the cost function is the Euclidean distance betwen the Mel frequency cepstral coefficients of each pair of frames. The optimization problem (6) can be solved using dynamic programming [12]. The warped template is given by

$$\hat{p}(n,k) = p_{\hat{l}}(\hat{t}(n),k), \quad \text{for } k = 1,\ldots,K. \quad (7)$$

Note that in dynamic time warping, it is customary to warp the time scales of both the recording and the template to find the closest match. Here, we warp the time scale of the template to match that of the recording.

3.2 Time-Frequency Masking

Because speech and other signals are sparse in the time-frequency domain, mixtures of several such sources can be effectively separated by assigning each time-frequency bin to a single source [13]. This process is known as time-frequency masking, and is often used in single-channel source separation. First, a mask is generated by comparing the power in the warped template to a threshold:

$$r(n,k) = \begin{cases} 1, & \text{if } |\hat{p}(n,k)|^2 > \gamma(k) \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

The threshold $\gamma(k)$ is a tuning parameter. In our experiments, we set it so that roughly 10% of the mask frames are 1 for each frequency bin k.

To isolate the keyword in the recording from interference and noise, we apply the time-frequency mask to each channel:

$$\hat{c}_m(n,k) = x_m(n,k) r(n,k). \quad (9)$$

If the signals are indeed sparse and if the mask is a good fit, then for nonzero values of $\hat{c}_m(n,k)$, we have $|a_m(k)s(n,k)|^2 \gg |z_m(n,k)|^2$, so that

$$\hat{\mathbf{c}}(n,k) \approx \mathbf{a}(k) s(n,k). \quad (10)$$

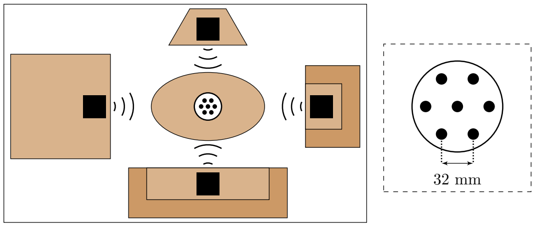

Fig. 3. Experimental setup using a MEMS microphone array in a living room.

3.3 Relative Transfer Functions

Finally, we use the masked signals to estimate the relative transfer functions. We compute the sample covariance matrix of the masked source spatial images:

$$\hat{\Sigma}_c(k) = \frac{1}{N}\sum_{n=1}^{N} \hat{\mathbf{c}}(n,k)\hat{\mathbf{c}}^H(n,k). \tag{11}$$

If (10) held exactly, then $\hat{\Sigma}_c(k)$ would be a rank-1 matrix proportional to $\mathbf{a}(k)\mathbf{a}^H(k)$. Let $\mathbf{u}(k)$ be the singular vector corresponding to the largest singular value of $\hat{\Sigma}_c(k)$. Then the estimated RTF vector is

$$\hat{\mathbf{a}}(k) = \frac{\mathbf{u}(k)}{u_1(k)}. \tag{12}$$

This is a special case of covariance whitening RTF estimation [6] where the noise is reduced by time-frequency masking rather than whitening. Related classification-based RTF estimation methods incorporate speech presence probabilities [14] and sparsity assumptions [15] to improve the mask.

4 Experiments

To evaluate the performance of the proposed separation method, we present empirical results for RTF estimation and source separation in a cocktail party scenario. The recording device, which is designed for voice assistant applications, is a circular array of $M = 7$ digital MEMS microphones spaced about 32 mm apart, as shown in Figure 3. The array sits on a coffee table in the center of a living room ($T_{60} \approx 400$ ms) and four signals are emitted from loudspeakers placed on a television stand, sofa, chair, and dining table between one and two meters away. One source is designated the target and the other three are interference.

Impulse responses were measured using sweep signals and used to simulate speech mixtures from prerecorded data. The keywords are taken from a crowdsourced database of one-second spoken commands [16]. The samples were recorded in widely varying environments with different equipment, reverberation characteristics, and noise levels, so the acoustic simulation is less realistic than

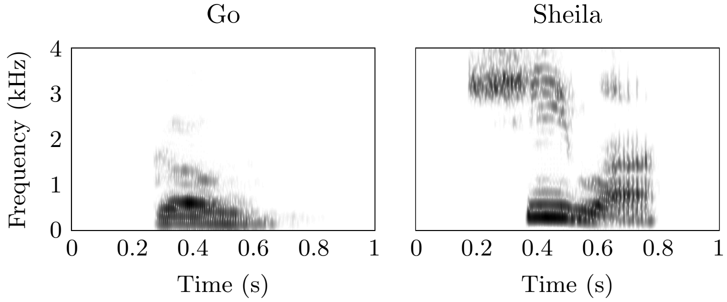

Fig. 4. Spectrograms of two keywords with different RTF estimation performance.

it would be with samples recorded in controlled anechoic conditions. Recordings with excessive background noise were removed and the clips were normalized to a constant average power. The experiments below use a set of $L = 500$ templates and a separate test set of 100 utterances for each keyword. For each trial, four ten-second speech clips are selected at random from a subset of the TIMIT database [17]. The mixtures also include a multichannel recording of living room background noise from appliances and ventilation. In these experiments, all signals are sampled at 8 kHz and the STFT uses a length-1024 discrete Fourier transform, a von Hann window of length 1024 samples (128 ms), and a step size of 256 samples (32 ms).

4.1 Relative Transfer Function Estimation Results

The MPDR beamformer, like many related multichannel filters, reduces noise and interference by projecting the mixture vector onto the RTF vector of the target source. If the estimated RTF vector is not parallel to the source image vector, the source will be distorted and unwanted noise might be amplified. Thus, to measure RTF estimation performance, we use the angle between the true and estimated RTF vectors, averaged across frequency bins:

$$\text{RTF Error} = \frac{1}{K}\sum_{k=0}^{K-1} \arccos \operatorname{Re}\left[\frac{\hat{\boldsymbol{a}}^H(k)\boldsymbol{a}(k)}{|\hat{\boldsymbol{a}}(k)|\,|\boldsymbol{a}(k)|}\right]. \tag{13}$$

Figure 5 shows RTF estimation error as a function of the input signal-to-interference-plus-noise ratio (SINR) of the keyword recording. The plots on the left show estimation performance using the ideal binary mask (IBM), which is one when $|s(n,k)|^2 > |z_1(n,k)|^2$ and zero otherwise. The IBM experiment shows the effect of keyword choice on RTF estimation performance if the keyword and noise signals were known perfectly. The plots on the right show the performance of the proposed method with template matching and mask estimation.

It is clear that longer keywords are better than shorter keywords, but there is significant variation even between keywords with the same number of syllables.

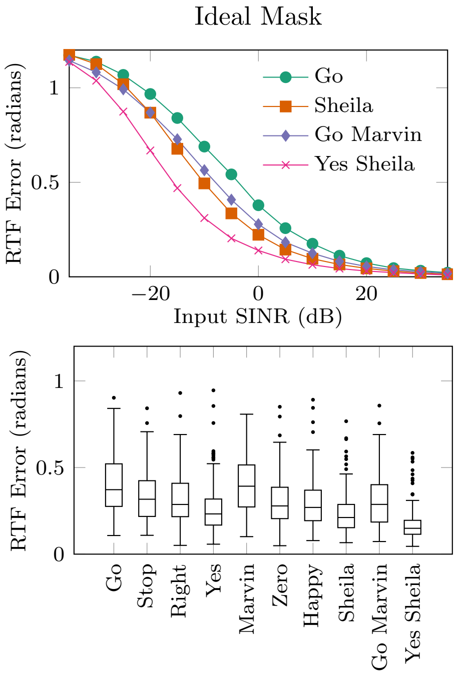

Fig. 5. RTF estimation performance using different keywords. Top: RTF error versus input SINR. Bottom: RTF error at 0 dB input SINR.

Keywords that contain sibilants ("yes", "Sheila") and thus strong high-frequency content appear to outperform keywords that do not. These keywords are easier to align with templates and cover more of the speech spectrum.

4.2 Source Separation Results

The ultimate goal of the proposed method is to improve source separation performance in a far-field speech capture system. We measure separation performance using the signal-to-error ratio (SER), computed in the time domain:

$$\mathrm{SER} = 10 \log_{10} \frac{\sum_t s^2(t)}{\sum_t \left(\hat{s}(t) - s(t)\right)^2}. \tag{14}$$

Figure 6 shows the SER for mixtures of four speech sources and background noise at an input SINR of about −4 dB. The plot on the left shows the SER as a function of the keyword input SINR (the input SINR of the unknown speech was not varied). The proposed method provides a roughly 20 dB keyword SINR improvement over the blind RTF estimator, which selects the dominant singular vector of $\hat{\Sigma}_x(k)$ at each frequency. There is a significant gap between the ideal and estimated mask performance, suggesting that there is room for improvement

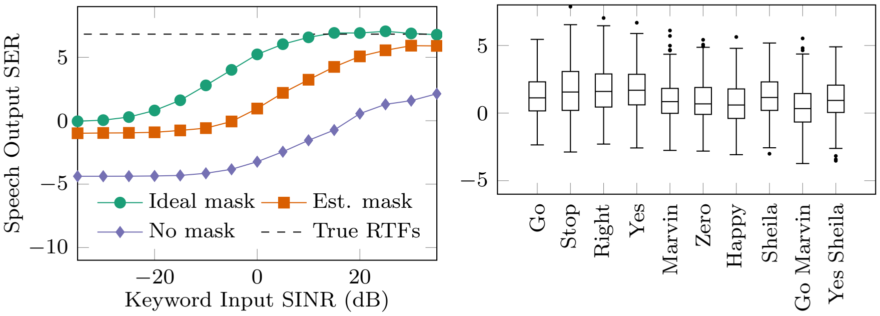

Fig. 6. Source separation performance with four speech sources. Left: Median speech output SER versus keyword input SINR with keyword "Yes Sheila". Right: Speech output SER at 0 dB keyword input SINR.

in the template-matching and mask estimation algorithms. The plot on the right shows the output SER when the keyword input SER is 0 dB. The output SERs vary less than the RTF errors for different keywords, and keywords that include sibilants do not have a clear advantage. Since the average spectrum of speech signals is concentrated at low frequencies, high-frequency RTF errors have a smaller impact on the separated speech signal.

5 Conclusions

The experiments show that speech keywords can be used as pilot signals to estimate the RTFs of a source in a noisy mixture. The proposed method is most useful when the interference and noise statistics are not known in advance, so covariance whitening and other model-based RTF estimation methods cannot be applied. In these situations, our experiments suggest that keyword-based RTF estimation can dramatically improve source separation performance.

The accuracy of the RTF estimate appears to depend on the length and the spectral content of the keyword. The most useful keywords have a variety of sounds, making them easy to separate by masking and ensuring that the full speech spectrum is captured by the template. The choice of keyword has a smaller impact on the performance of the separator, suggesting that the method may be useful for some applications even with keywords that are short and spectrally concentrated. In this work, we have used relatively simple algorithms for template matching, mask estimation, and source separation. While these are adequate for this proof of concept, our results suggest that more sophisticated algorithms could improve performance.

Many source separation methods rely on assumptions about the geometry of the array or the statistics of the source signals. However, we can also leverage information about the *content* of the signals. This study has shown that we can effectively separate a speech source from strong interference based only on our knowledge of a single word.

References

1. Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A.: A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(4) (2017) 692–730
2. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2013) 126–130
3. Doclo, S., Kellermann, W., Makino, S., Nordholm, S.E.: Multichannel signal enhancement algorithms for assisted listening devices. IEEE Signal Processing Magazine **32**(2) (2015) 18–30
4. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Transactions on Signal Processing **49**(8) (2001) 1614–1626
5. Cohen, I.: Relative transfer function identification using speech signals. IEEE Transactions on Speech and Audio Processing **12**(5) (2004) 451–459
6. Markovich, S., Gannot, S., Cohen, I.: Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Transactions on Audio, Speech, and Language Processing **17**(6) (2009) 1071–1086
7. Makino, S., Lee, T.W., Sawada, H.: Blind speech separation. Springer (2007)
8. Pedersen, M., Larsen, J., Kjems, U., Parra, L.: Convolutive blind source separation methods. In: Springer Handbook of Speech Processing. Springer (2008) 1065–1094
9. Corey, R., Singer, A.: Real-world evaluation of multichannel audio enhancement using acoustic pilot signals. In: Asilomar Conference on Signals, Systens, and Computers. (2017)
10. Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2014) 4087–4091
11. Van Trees, H.: Optimum array processing. Wiley (2004)
12. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing **26**(1) (1978) 43–49
13. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing **52**(7) (2004) 1830–1847
14. Araki, S., Okada, M., Higuchi, T., Ogawa, A., Nakatani, T.: Spatial correlation model based observation vector clustering and mvdr beamforming for meeting recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2016) 385–389
15. Koldovskỳ, Z., Málek, J., Gannot, S.: Spatial source subtraction based on incomplete measurements of relative transfer function. IEEE Transactions on Audio, Speech, and Language Processing **23**(8) (2015) 1335–1347
16. Warden, P.: Speech commands: A public dataset for single-word speech recognition. Dataset available from `http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz` (2017)
17. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D.: TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Web Download (1993)