Dictionary-Based Sparse Reconstruction of Incomplete Relative Transfer Functions

Zbyněk Koldovský* and Sharon Gannot†
*Acoustic Signal Analysis and Processing Group, Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Czech Republic.
†The Alexander Kofkin Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel.

***Abstract*— For estimating the relative transfer function (RTF) of a speaker from noisy multi-microphone recordings, several statistical methods have been proposed. The estimation accuracy is different over frequencies, which mostly depends on the frequency-dependent signal-to-noise ratio (SNR). Provided that the low-SNR frequencies are identified, the corresponding values of the estimated RTF can be replaced through interpolation using the frequencies with high SNR. In this study, we explore interpolation techniques based on the sparse reconstruction of an incomplete RTF which is obtained when low-SNR values are neglected. Compared to previous attempts where the approximate sparsity of the time-domain representation of RTF (relative impulse response) is exploited, in this paper, we use learned sparse dictionaries trained on dense measurements of RTFs within a confined area of the target speaker. These measurements are obtained from the recently released MIRaGe database acquired in a real room.**

***Index Terms*— Room Impulse Responses, Relative Transfer Function, Sparse Representations, Sparse Dictionaries, Dictionary Learning**

I. Introduction and Problem Formulation

Consider a speaker recorded in reverberant environment using several microphones. The signal observed by the ith microphone is described as [1]

$$x_i(n) = \{h_i \star s\}(n) + y_i(n), \tag{1}$$

in the time-domain, and approximately

$$X_i(k,\ell) = H_i(k)S(k,\ell) + Y_i(k,\ell), \tag{2}$$

in the time-frequency domain. Here, $\star$ denotes the convolution operator, and n, k, and ℓ stand for the sample, frequency, and frame index, respectively. The speaker's voice is represented by s and S, while y_i and Y_i contain the other interfering and noise signals. h_i is the room impulse response (RIR) characterizing the acoustic path between the speaker and the ith microphone, and H_i is its frequency domain representation, i.e., the acoustic transfer function (ATF).

By considering a pair of microphones i and j, $i \neq j$, the observed signals can be described in a relative way as

$$x_i(n) = s_i(n) + y_i(n), \tag{3}$$

$$x_j(n) = \{g_{ij} \star s_i\}(n) + y_j(n), \tag{4}$$

This work was supported by The Czech Science Foundation through Project No. 20-17720S.

or

$$X_i(k,\ell) = S_i(k,\ell) + Y_i(k,\ell), \tag{5}$$

$$X_j(k,\ell) = G_{ij}(k)S_i(k,\ell) + Y_j(k,\ell), \tag{6}$$

where $s_i = \{h_i \star s\}$ and $S_i = H_i(k)S(k,\ell)$ denote the speakers signals as received on the ith microphone (also referred to as the ith image of s), and g_{ij} is the relative impulse response (ReIR) characterizing the relative difference between s_i and s_j, related to the ith microphone. The respective frequency-domain counterpart of g_{ij} is G_{ij} called the relative transfer function (RTF) [2].

When the above transfer functions (or respective impulse responses) are known, various spatial processors such as MPDR, MVDR, and LCMV can be applied in order to separate the desired speech from the noisy recordings [3], [4]. The advantage of RTFs (ReIRs) is that they can be consistently estimated from the microphone observations. Conventional least-squares estimators (either time or frequency domain) can be used when noise-free recordings are available ($y_i = y_j = 0$). When the noise is isotropic and stationary, a model-based estimator from [5] can be applied to noisy recordings. When the covariance of noise is known, the generalized eigenvalue decomposition (GEVD) can be applied [6]. Some situations with interfering directional sources can be handled using Blind Source Separation (BSS) [7], [8], [9]. More recently, deep learning-based approaches have been proposed in order to classify time-frequency points according to the SNR [10], which can be further used for the RTF estimation [11], [12].

The RTF estimators achieve different accuracy per frequency channel, which mostly depends on the level of mismatch between the assumed model or training set and the observed data. For example, conventional least-squares estimators are sensitive to the frequency-dependent signal-to-interference-plus-noise ratio (SINR). By assuming that knowledge about the accuracy is available, the values of the estimator of $G_{ij}(k)$, denoted as $\hat{G}_{ij}(k)$, can be labeled either as good or bad (or weighted according to a level of uncertainty). The labeled estimator could be treated as *incomplete*, meaning that the inaccurate values of $\hat{G}_{ij}(k)$ are neglected or outweighted. Then, an improved RTF estimate can obtained through reconstructing the incomplete RTF in a sparse domain [13].

The success of this approach depends on several factors where the most important one resides in the knowledge of an

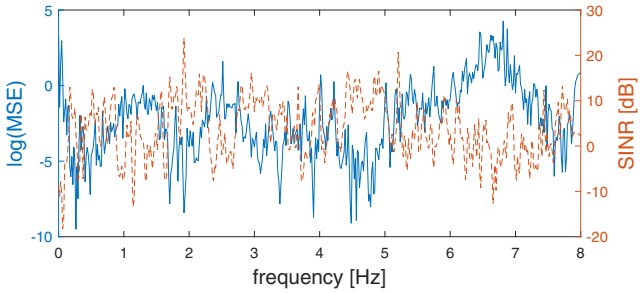

Fig. 1. Mean square error (MSE) between the RTFs estimated from clean and noisy recording (the left y-axis) and the SINR (the right y-axis) as functions of frequency.

appropriate sparsity domain. That is, the domain in which the RTFs (respectively, the ReIRs) have sparse representations is of primal interest. In our previous works such as [13], [14], the time-domain was considered, which comes from the fact that ReIRs, i.e. the time-domain representations of RTFs, are approximately sparse (namely, have many coefficients that are close to zero).

However, this hypothesis is not very accurate, especially, in highly reverberated environments where the ReIRs have long tails. To obtain a better sparsity domain, a learning-based approach can be used in order to find a sparse dictionary that is useful in the given acoustic environment (and likely not elsewhere) [15].

This work is focused on learning such dictionaries. In the following section, the problem of sparse reconstruction of incomplete RTFs using a dictionary is formulated. Section 3 describes our practical approach to learn sparse dictionaries of this purpose using the standard K-SVD algorithm [16]. Section 4 reports results of an experimental study carried out using the recently released MIRaGe dataset that provides detailed measurements of ReIRs in a real room [17]. Conclusions are drawn in Section 5.

II. Sparse Reconstruction of Incomplete RTFs

Let us begin with an example where a speaker is recorded by two microphones from a distance of 1 m; the microphone inter-distance is 8 cm; the recording is 2 s long; the sampling frequency is 16 kHz. Using the clean speech recording and the least-squares procedure, the ReIR of length 1024 taps is estimated. The Discrete Fourier transform (DFT) of the ReIR is considered as the ground truth RTF. Then, another speaker and bubble noise are added to the recording so that the global SINR is 5 dB. The ReIR is estimated using least-squares from the noisy data, and the corresponding RTF is compared with the ground truth RTF.

Figure 1 depicts a joint plot of the frequency-dependent mean square error (MSE) between the RTFs (the left y-axis) and the frequency-dependent SINR (the right y-axis). The figure demonstrates a clear correspondence between the SINR and the accuracy of the least-squares RTF estimator. When the SINR is sufficiently high, the MSE tends to be small, and vice verse.

The example motivates us for interpreting the noisy RTF estimator as an incomplete RTF measurement. We will work with the assumption that information about SINR is available so that the set of frequencies for which the SINR is "sufficiently" high is known; let the set of such frequencies be denoted by $\mathcal{S}$. We admit that such information provides a strong knowledge and might be difficult to obtain. Nevertheless, the purpose of this work is to focus on the other important problem: Can we infer the unknown values of the incomplete RTF (iRTF)?

Let $\mathbf{D}$ be an $L \times M$ matrix representing a dictionary, where L corresponds to the length of ReIRs which is equal to the DFT resolution, and M is the number of atoms in the dictionary. When $\mathbf{D}$ is an appropriate sparse dictionary, the iRTF can be completed through finding its sparsest representation in $\mathbf{D}$, that is, by solving

$$\min_{\mathbf{x}\in\mathbb{R}^M} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{F}_{k,:}\mathbf{D}\mathbf{x} = \hat{G}_{ij}(k), \qquad k \in \mathcal{S}, \tag{7}$$

where $\mathbf{F}$ denotes the $L \times L$ DFT matrix, $\mathbf{F}_{k,:}$ denotes its kth row, and $\|\mathbf{x}\|_0$ is equal to the number of nonzero elements in $\mathbf{x}$. Once such $\mathbf{x}$ is found, the reconstructed ReIR is equal to $\mathbf{D}\mathbf{x}$.

The complexity of (7) is known to be NP-hard, therefore, several convex relaxations are considered in the literature [18] such as basis pursuit, basis pursuit denoising (BPDN), and LASSO. In this work, we consider the following convex program

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^M} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad & |\Re\{\mathbf{F}_{k,:}\mathbf{D}\mathbf{x} - \hat{G}_{ij}(k)\}| \le \delta, \qquad k \in \mathcal{S}, \\ & |\Im\{\mathbf{F}_{k,:}\mathbf{D}\mathbf{x} - \hat{G}_{ij}(k)\}| \le \delta, \qquad k \in \mathcal{S}, \end{aligned} \tag{8}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real and imaginary parts of the argument, $\|\cdot\|_1$ denotes the ℓ_1 norm, and $\delta \ge 0$ is a threshold parameter.

The latter optimization problem corresponds to ℓ_1-minimization with ℓ_∞ constraints. For $\delta = 0$, it corresponds with the basis pursuit formulation, and it can be reformulated as a (real-valued) linear programming problem for every $\delta \ge 0$ [19]. Compared to BPDN or LASSO, the constraints in (8) guarantee that the values of the reconstructed RTF do not differ from those of the incomplete one by more that δ, simultaneously for every $k \in \mathcal{S}$. This provides a more controlled solution than with LASSO [20].

III. Dictionary Learning

The choice of the dictionary $\mathbf{D}$ plays a crucial role in sparse reconstruction. In our problem, every RTF that should be reconstructed from iRTF should have a sufficiently sparse representation in $\mathbf{D}$. A general-purpose $\mathbf{D}$ is hard to find and, most likely, such dictionary does not exist.

In previous works, the typical time-domain structures of ReIRs, i.e. their approximate sparsity due to exponential decay, have been exploited [13]. This corresponds with the choice $\mathbf{D} = \mathbf{I}_L$, where $\mathbf{I}_L$ is the $L \times L$ identity matrix. Each atom (column of $\mathbf{D}$) represents an integer delay filter, so ReIRs

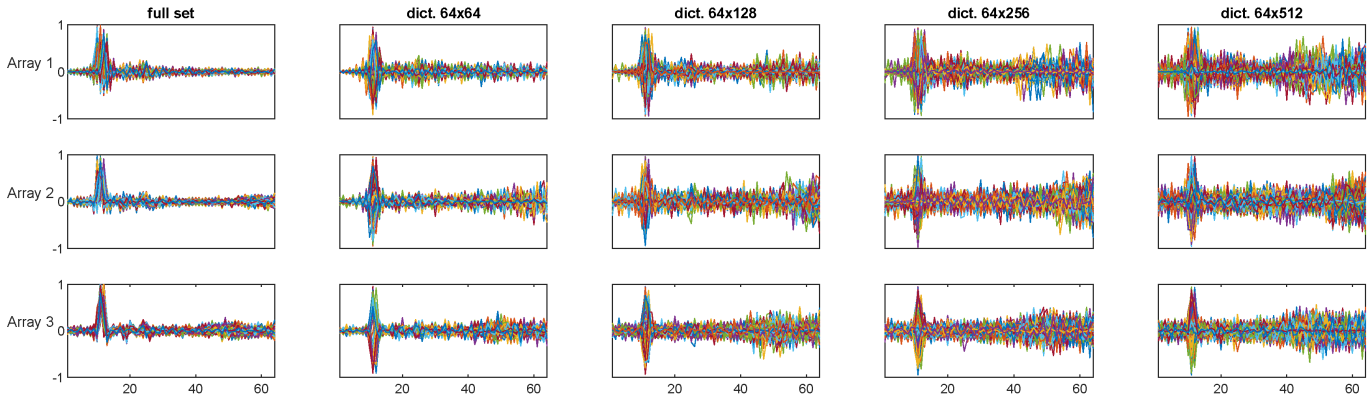

Fig. 2. Full sets of 16416 ReIRs (all target positions within the volume, 4 microphone pairs - the central mic being the reference one) and the atoms of dictionaries with 64, 128, 256, and 512 atoms (each of length 64) derived from the full sets using K-SVD (10^4 iterations starting from the identity matrix). The global delay (anti-causal part) of the ReIRs is 10 taps.

are assumed to be sparse linear combinations of the delays. A more advanced approach is when $\mathbf{D}$ involves also some fractional-delay filters (so-called "oversampled domain") or even corresponds to the continuous dictionary involving all fractional-delay filters [14].

However, the efficiency of the time-domain dictionaries is limited, especially, due to the reverberation (most of the ReIRs' coefficients are nonzero). More efficient dictionaries tailored to the particular acoustic scenario might be learned from on-site measurements. Such measurements must involve a sufficient number of ReIRs corresponding to potential positions of the target source. In this paper, we conduct such experiment with the recently released MIRaGe database.[1]

MIRaGe contains measurements of excitation signals played by a loudspeaker from positions that form a dense grid within a $46 \times 36 \times 32$ cm volume (the speaker's area). The setup is situated in an acoustic laboratory which is a $6 \times 6 \times 2.4$ m rectangular room with variable reverberation time. Three reverberation levels with T_{60} equal to 100, 300, and 600 ms are provided. The speaker's area involves 4104 positions which form the cube-shaped grid with spacing of 2-by-2 cm over the x and y axes and 4 cm over the z axis. Also, MIRaGe contains a complementary set of measurements that provide information about the positions placed around the room perimeter with spacing of $\approx$1 m, at a distance of 1 m from the wall. These positions are referred to as the out-of-grid positions (OOG) and can be used for simulating interfering sources. All measurements were recorded by six static linear microphone arrays (5 mics per array with the inter-microphone spacing of -13, -5, 0, $+5$ and $+13$ cm relative to the central microphone); for all details about the database, see [17].

For our experiment, we use the white noise excitation signals to compute ReIRs of length 8192 taps using least-squares; the anti-causal part, referred to as global delay, has 128 taps. The ReIRs are then truncated to the selected length L and global delay 10. In the experiments, we consider Arrays 1 through 3, and the reference microphone is the central microphone in each array. Hence, we have $4 \times 4104 = 16416$ ReIRs for each array and for each T_{60} setting to learn the dictionary.

As the learning method, we use the standard K-SVD algorithm [16]. The ReIRs are truncated so that $L = 64$ and the global delay is 10 taps. The size of the dictionary is $M = 64$, 128 ,256, and 512, respectively. The number of K-SVD iterations is set to 10^4.

The resulting dictionaries for Arrays 1 through 3 when $T_{60} = 100$ ms are visualized in Fig. 2. In the first column, the figures show all ReIRs in the corresponding training set. The main peaks (direct-path delay) of all ReIRs are concentrated around tap 10, which is the global delay. For Array 1, the position of the main peak is within taps 8-16 while, for Arrays 2 and 3, it is 9-14 and 9-13, respectively. This is due to the growing distance of Arrays 1 through 3, hence, the decreasing range of the angle of arrival. Also, the early reflection and reverberation tails of the ReIRs are more intensive for more distant arrays. The atoms of the dictionaries with 64, 128, 256, and 512 atoms are shown in columns 2 through 5. For small M, the atoms mainly involve the main peaks of the ReIRs around tap 10. However, with growing M, there is an increasing number of atoms having a more complicated structure, probably involving early reflections and reverberation tails of the ReIRs.

IV. Experimental Results

A. Experimental Setup

The experimental scenarios were realized using the MIRaGe database. For a given array of microphones, T_{60} setting, and dictionary $\mathbf{D}$, we carry out the following experiment: The target (female) speech of length 10 s is played from a selected position within the target volume using RIRs of length 8192 taps. Similarly, an interfering speaker (male) is simulated from the OOG position #24. The spatial images of the speakers' utterances are summed together with a babble noise sequence. The initial SINR computed over the entire sequence is nearly

[1] https://asap.ite.tul.cz/downloads/mirage/

0 dB, depending on the target speaker position, microphone, array, and T_{60}.

Then, a 2 s long interval of the mixed signals is used to estimate the RTFs of microphones #1, 2, 4, and 5 with respect to microphone #3 of each array with global delay 10; the time-domain least-squares approach is used; the estimated ReIR is transformed using FFT to obtain the estimated RTF. As shown in Fig. 1, the accuracy of this estimator highly depends on the frequency-dependent SINR (good in frequencies where the target source is dominant; very poor where the SINR is small).

To define the iRTF, $\mathcal{S}$ is chosen according to the known frequency-dependent SINR. Namely, $\mathcal{S}$ consists of p percent of the frequency bins with the highest SINR. Then, the new RTF estimate is reconstructed from the iRTF by solving (8). The result is compared with the ground-truth RTF in terms of the mean-square error (MSE).

The following approaches and estimators are compared: DICT and TIME correspond to the noisy estimate of RTF reconstructed, respectively, in the dictionary and time domain. Oracle variants of these methods perform the reconstruction using the same set $\mathcal{S}$, however, the iRTF contains coefficients of the ground-truth RTF. Next, NearestNeighbour is a brute-force approach that utilizes the full set of ground-truth RTFs as prior knowledge. It outputs the RTF from the full set whose coefficients, restricted to $\mathcal{S}$, are closest to the noisy-estimated iRTF in the mean square sense. Finally, CENTRAL stands for a naïve approach that always gives the ground-truth RTF corresponding to the central position of the grid; RANDOM takes the ground-truth RTF for a random position within the grid.

In the following, we show and discuss results of the experiments under various settings. The results are averaged over all possible 4104 positions of the target source within the grid, over microphones #1, 2, 4 and 5. In case of DICT and TIME, the iRTF is estimated using 2 s long intervals of noisy signals, i.e., there are 5 different estimates from the 10s long signals; the average is taken also over these estimates.

In particular, we focus on the averaged MSE as a function of the percentage of frequency bins p included in $\mathcal{S}$. For p close to 100%, DICT and TIME approach the original RTF estimates obtained from the noisy signals. The estimators coincide when $p = 100\%$ and $\delta = 0$. Similarly, oracle DICT and oracle TIME approach the ground true RTF when p is close to 100%, so they achieve MSE that is close to 0.

B. Results

Figure 3 shows results for Array 1, $T_{60} = 100$ ms, and $M = 512$. Oracle DICT yields a steeply decreasing MSE with growing p, and the decrease is significantly steeper than that of Oracle TIME. This shows that the dictionary has been trained well because every RTF can be reconstructed accurately knowing very few values. While this gives the proof of concept, the behavior of DICT and TIME shows the practical utility as they are using the values of the noisy RTF estimate.

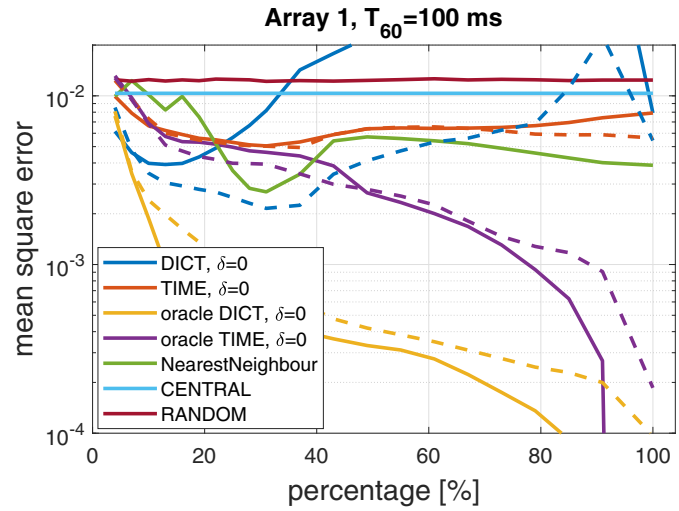

Fig. 3. Comparison of methods in terms of the averaged mean square error. Solid lines correspond to $\delta = 0$; dashed lines correspond to $\delta = 0.1$.

The MSE by DICT and TIME is minimal when p is between 10 and 30%, depending on the approach and δ. These values correspond with the approximate number of frequency bins where the target speech is dominant and where the RTF estimate is accurate enough. The lowest (best) MSE by DICT is smaller than that of TIME, which points to the efficiency of the dictionary. For the higher values of p, the iRTF is contaminated by significantly biased estimates of the RTF, which causes the growth of the MSE. DICT with $\delta = 0$ appears to be very sensitive to such values, nevertheless, it is quite robust when $\delta = 0.1$.

Finally, the naïve approaches CENTRAL and RANDOM do not bring any quality to the RTF estimation compared to the other methods. The brute-force NearestNeighbour yields good performance when $p = 30\%$ and shows a robustness when p is higher. On the other hand, it yields large MSE for the critically low values of $p < 20\%$, which is of greater interest.

Figure 4 compares DICT and TIME and their oracle variants when T_{60} is varying; the array index is 1; $\delta = 0.1$; $M = 512$. Note that with $\delta = 0.1$, oracle methods do not achieve zero MSE as $p = 100\%$, because the solution of (8) slightly deviates from the ground-truth RTF. The behaviour of DICT and TIME is similar to that in Fig. 3, and, as could be expected, the best achieved MSE tends to be higher (worse) as T_{60} grows. Interestingly, the oracle methods exhibit the opposite behavior: For a given value of p, the MSE tends to be lower when T_{60} is higher.

Figure 5 compares the same methods when the distance between the speaker and microphone array varies from 1 to 3 m; $T_{60} = 300$ ms, $M = 512$. The methods behave consistently with the results in Figure 3; the minimum MSE achieved by DICT and TIME is higher when $T_{60} = 300$ ms. The compared methods yield overall higher MSE when the array-source distance is 3 m (Array 3) than when it is 1 or 2 m.

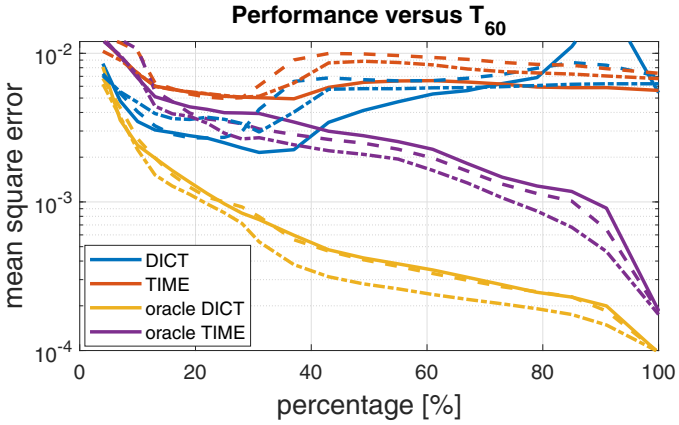

Fig. 4. Mean square error versus percentage shown for different T_{60} settings (Array 1). Solid line: $T_{60} = 100$ ms, dashed lines: $T_{60} = 300$ ms, and dashdotted lines: $T_{60} = 600$ ms; computed with $\delta = 0.1$.

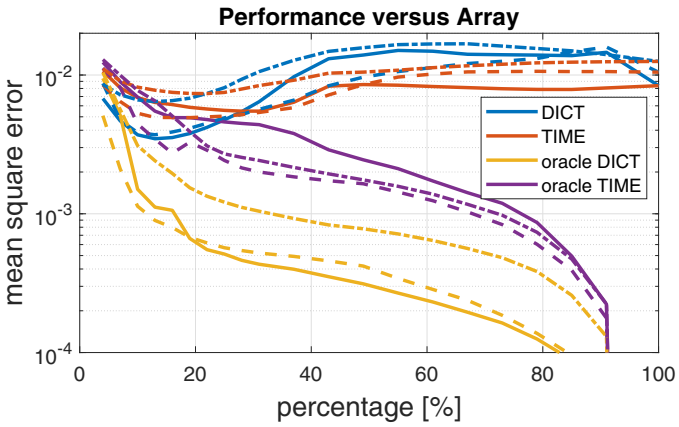

Fig. 5. Mean square error as a function of percentage for different distances of the microphone array (array index equals the distance in meters) for $T_{60} =$ 300 ms. Solid line: Array 1, dashed lines: Array 2, and dashdotted lines: Array 3.

V. Conclusions

We have shown that a sparsity domain of RTFs can be learned for a specific room, target source area, and a microphone array position. Provided that high-SNR frequencies can be identified, conventional RTF estimators can be used to compute an incomplete RTF estimate. A complete RTF estimate can be obtained through finding the sparsest representation of the incomplete RTF in the sparsity domain. The experiments have shown that this approach can yield significantly more accurate RTF estimates than the conventional methods when the RTF is estimated from noisy recordings.

References

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, Wiley Publishing, 1st edition, 2018.

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

[3] Harry L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, John Wiley & Sons, Inc., 2002.

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

[5] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE transactions on signal processing*, vol. 44, no. 8, pp. 2055–2063, 1996.

[6] Shmulik Markovich, Sharon Gannot, and Israel Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Tran. on Au., Sp., and Lang. Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[7] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*, Springer, 2007.

[8] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined lcmv-trinicon beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 320–332, 2017.

[9] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1050–1064, Feb 2019.

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

[11] S. E. Chazan, J. Goldberger, and S. Gannot, "Dnn-based concurrent speakers detector and its application to speaker extraction with lcmv beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6712–6716.

[12] Jiří Málek, Zbyněk Koldovský, and Marek Boháč, "Block-online multi-channel speech enhancement using dnn-supported relative transfer function estimates," *IET Signal Processing*, vol. 14, pp. 124–133, 2020.

[13] Z. Koldovský, J. Málek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1335–1347, Aug 2015.

[14] Z. Koldovský and P. Tichavský, "Sparse reconstruction of incomplete relative transfer function: Discrete and continuous time domain," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 394–398.

[15] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[16] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[17] Jaroslav Čmejla, Tomáš Kounovský, Sharon Gannot, Zbyněk Koldovský, and Pinchas Tandeitnik, "Mirage: Multichannel database of room impulse responses measured on high-resolution cube-shaped grid in multiple acoustic conditions," in *Proceedings of European Signal Processing Conference*, Jan. 2020, pp. 56–60.

[18] David L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[19] C. Brauer, D. A. Lorenz, and A. M. Tillmann, "A primal-dual homotopy algorithm for ℓ_1-minimization with ℓ_∞-constraints," *Computational Optimization and Applications*, vol. 70, no. 2, pp. 443–478, 2018.

[20] P. Rajmic, Z. Koldovský, and M. Daňková, "Fast reconstruction of sparse relative impulse responses via second-order cone programming," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 364–368.