MULTISOURCE MINT USING CONVOLUTIVE TRANSFER FUNCTION

Xiaofei Li[1], *Sharon Gannot*[2], *Laurent Girin*[1,3] *and Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes, France
[2]Faculty of Engineering, Bar-Ilan University, Israel
[3]Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab, France

ABSTRACT

The multichannel inverse filtering method, i.e. multiple input/output inverse theorem (MINT), is widely used. However, it is usually performed in the time domain, and based on the long room impulse responses, thus it has a high computational complexity and a large number of near-common zeros. In this paper, we propose to perform MINT in the short-time Fourier transform (STFT) domain, in which the time-domain filter is approximated by the convolutive transfer function. The oversampled STFT is used to avoid frequency aliasing, which however leads to a common zero region in the subband frequency response due to the frequency response of the STFT window. A new inverse filtering target function concerning the STFT window is proposed to overcome this problem. In addition, unlike most studies using MINT for single source dereverberation, the multisource MINT is proposed for both source separation and dereverberation.

Index Terms— Multisource MINT, CTF

1. INTRODUCTION

Multichannel inverse filtering (or multichannel equalization) of room acoustic aims at recovering the source signal from the convolutive recording signals. To this aim, the multiple-input/output inverse theorem (MINT) method was first proposed in [1]: An inverse filter is estimated with respect to the known room impulse responses (RIR), and applied to the microphone signals, preserving the desired source and suppressing the interfering sources.

It is known that MINT is sensitive to RIR perturbations (misalignment / estimation error) and to microphone noise. To improve the robustness of MINT to the RIR perturbations, many techniques have been proposed, preserving not only the direct-path impulse response but also the early reflections, such as channel shortening [2], infinity- and p-norm optimization-based channel shortening/reshaping [3], partial MINT [4, 5], etc. In addition, the energy of the inverse filter was used in [6] as a regularization term to avoid the amplification of filter perturbations and microphone noise.

The above techniques were introduced in the time domain. Time-domain RIRs are usually very long, which leads to a high computational complexity and a large number of near-common zeros among microphones. To shorten the room filters, several variant of subband MINT were proposed [7, 8, 9, 10, 11]. The key issues in the filter-bank design are 1) the time-domain RIRs should be well approximated in the subband domain, 2) the frequency response of each filter-bank should be fully excited, i.e. should not involve the frequency components with the magnitude close to zero. Otherwise, these components are common to all channels, and are problematic in the MINT application. To satisfy the second condition, the filter-bank either is critically sampled [7, 8], which suffers from frequency aliasing, or has a flat-top frequency response [9, 10, 11], which may suffer from time aliasing.

In this paper, we propose a subband MINT based on the widely-used short-time Fourier transform (STFT). The time-domain RIR can be exactly represented in the STFT domain by the cross-band filters [12], and further approximated by its band-to-band version, aka the convolutive transfer function (CTF) [13, 14]. Based on CTF, a Lasso method and an Expectation-Maximization method were respectively proposed in [15] and [16] for source separation. A Hamming window with 75%-overlap is used in this work, which avoids both frequency aliasing and time aliasing. However, the Hamming window is not flat-top, namely there is a large region close to the margin of the main lobe having a magnitude close to zero. To overcome this problem, instead of using the conventional impulse function as the target of the inverse filtering, we propose a new target, which has a frequency response corresponding to the STFT window. In addition, all the above-mentioned techniques were proposed for single source dereverberation. To our knowledge, the multisource case has been rarely studied, even if the multisource MINT was presented in the original paper [1]. In this paper, a CTF-based multisource MINT method is proposed for both source separation and dereverberation.

The rest of this paper is organized as follows. Section 2 presents the CTF model. The proposed multisource MINT

This research has received funding from the ERC Advanced Grant VHIA (#340113).

method is described in Section 3. Experiments are presented in Section 4. Section 5 concludes the paper.

2. CONVOLUTIVE TRANSFER FUNCTION

Let us first consider a noise-free single-microphone signal $x(n)$ being the result of the convolution of a single source signal $s(n)$ with a RIR $a(n)$: $x(n) = a(n) \star s(n)$. Let $x_{p,k}$ denote the STFT coefficients of $x(n)$, where p and k denote the frame index and the frequency index, respectively. The cross-band filter model consists in representing the STFT coefficient $x_{p,k}$ as a summation over multiple convolutions (between the STFT-domain source signal and filter) across frequency bins:

$$x_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} s_{p-p',k'} \; a_{p',k,k'}, \tag{1}$$

where N is the STFT window size. Let D denote the STFT frame step. If $D < N$, then $a_{p',k,k'}$ is non-causal, with $\lceil N/D \rceil - 1$ non-causal coefficients [12]. The number of causal filter coefficients is related to the reverberation time. For notational simplicity, let the filter index p' be in $[0, L_a - 1]$, with L_a being the filter length, i.e. the non-causal coefficients are shifted to the causal part, which only leads to a constant shift of the frame index of the source signal. Let $\tilde{w}(n)$ and $w(n)$ denote the STFT analysis window and synthesis window, respectively. The STFT-domain impulse response $a_{p',k,k'}$ is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k,k'} = (a(n) \star \zeta_{k,k'}(n))|_{n=p'D}, \tag{2}$$

which represents the convolution with respect to the time index n evaluated at frame steps, with

$$\zeta_{k,k'}(n) = e^{j\frac{2\pi}{N}k'n} \sum_{m=-\infty}^{+\infty} \tilde{w}(m)\, w(n+m)\, e^{-j\frac{2\pi}{N}m(k-k')}.$$

To simplify the analysis, we consider the CTF approximation, i.e., only band-to-band filters with $k = k'$ are considered:

$$x_{p,k} \approx \sum\nolimits_{p'=0}^{L_a-1} s_{p-p',k} a_{p',k} = s_{p,k} \star a_{p,k}. \tag{3}$$

3. MULTISOURCE CTF-BASED MINT

Based on the CTF approximation, the STFT domain convolutive mixture model with J sources and I microphones is

$$x^i_{p,k} = \sum_{j=1}^{J} a^{i,j}_{p,k} \star s^j_{p,k} + e^i_{p,k}, \tag{4}$$

where $e^i_{p,k}$ is the noise signal. The CTF $a^{i,j}_{p,k}$ is relating the j-th source to the i-th microphone. Let $p = 0, \ldots, P-1$ denote the frame index of the microphone signals, and as mentioned above, $p = 0, \ldots, L_a - 1$ denote the frame index of the CTFs. Since the proposed method is applied frequency-wise, hereafter the frequency index k is omitted unless necessary.

3.1. Problem Formulation for Inverse Filtering

Define the "CTF-domain" inverse filters as h^i_p with $i = 1, \ldots, I$ and $p = 0, \ldots, L_h - 1$, where L_h denotes the length of the inverse filters (identical for all i). The output of the inverse filtering is

$$\begin{aligned} y_p &= \sum_{i=1}^{I} h^i_p \star x^i_p \\ &= \sum_{j=1}^{J} s^j_p \star (\sum_{i=1}^{I} h^i_p \star a^{i,j}_p) + \sum_{i=1}^{I} h^i_p \star e^i_p, \end{aligned} \tag{5}$$

which comprises the mixture of the inverse filtered sources and the inverse filtered noise.

To facilitate the analysis, we denote the convolution in vector form. Define the convolution matrix for the CTF $a^{i,j}_p$ as

$$\mathbf{A}^{i,j} = \begin{bmatrix} a^{i,j}_0 & 0 & \cdots & 0 \\ a^{i,j}_1 & a^{i,j}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a^{i,j}_{L_a-1} & \ddots & \ddots & 0 \\ 0 & a^{i,j}_{L_a-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a^{i,j}_{L_a-1} \end{bmatrix} \in \mathbb{C}^{(L_a+L_h-1)\times L_h}, \tag{6}$$

and the vector of filter h^i_p as $\mathbf{h}^i = [h^i_0, \ldots, h^i_p, \ldots, h^i_{L_h-1}]^\top$, where $^\top$ denotes the transpose of a vector or a matrix. Then the convolution can be written as $h^i_p \star a^{i,j}_p = \mathbf{A}^{i,j}\mathbf{h}^i$.

3.2. Multisource MINT

To preserve a desired source, e.g. the j_d-th source, the inverse filtering of the CTF filters, i.e. $\sum_{i=1}^{I} \mathbf{A}^{i,j_d}\mathbf{h}^i$, generally should target to an impulse function d_p with the length of $L_a + L_h - 1$. To suppress the interfering sources, the inverse filtering of the CTF filters of the other sources, i.e. $\sum_{i=1}^{I} \mathbf{A}^{i,j\neq j_d}\mathbf{h}^i$, should target to a zero signal. Let $\mathbf{d}$ denote the vector form of d_p, and $\mathbf{0}$ denote a $(L_a + L_h - 1)$-dimensional zero vector. We define the following I-input

J-output MINT equation in the CTF representation (or, say, STFT-domain):

$$\begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{d} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{1,1} & \dots & \mathbf{A}^{I,1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^{1,j_d-1} & \dots & \mathbf{A}^{I,j_d-1} \\ \mathbf{A}^{1,j_d} & \dots & \mathbf{A}^{I,j_d} \\ \mathbf{A}^{1,j_d+1} & \dots & \mathbf{A}^{I,j_d+1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^{1,J} & \dots & \mathbf{A}^{I,J} \end{bmatrix} \begin{bmatrix} \mathbf{h}^1 \\ \vdots \\ \mathbf{h}^I \end{bmatrix}$$

or

$$\mathbf{g} = \mathbf{A}\mathbf{h}. \tag{7}$$

When the matrix $\mathbf{A} \in \mathbb{C}^{J(L_a+L_h-1)\times IL_h}$ is square or wide, namely $IL_h \geq J(L_a+L_h-1)$ and thus $L_h \geq \frac{J(L_a-1)}{I-J}$, (7) has an exact solution, which means an exact inverse filtering can be achieved. This condition implies an overdetermined recording system, i.e. $I > J$.

From [1], the solvable condition of (7) is that the CTFs of the desired source $a_p^{i,j_d}, i = 1, \dots, I$ do not have any common zero. The subband filters, i.e. CTFs, are much shorter than the time domain filters, and are thus likely to have much less near-common zeros, which is a major benefit. Unfortunately, the filter banks induced from the short-time windows will lead to some structured common zeros. From (2), for any RIR $a^{i,j}(n)$, its CTF (with $k' = k$) is computed as

$$a_{p,k}^{i,j} = (a^{i,j}(n) \star \zeta_k(n))|_{n=pD}, \tag{8}$$

with $\zeta_k(n) = e^{j\frac{2\pi}{N}kn} \sum_{m=-\infty}^{+\infty} \tilde{w}(m)\, w(n+m)$ being the cross-correlation of the analysis window $\tilde{w}(n)$ and the synthesis window $w(n)$ modulated (frequency shifted) by $e^{j\frac{2\pi}{N}kn}$. This cross-correlation has a similar frequency response as the windows $\tilde{w}(n)$ and $w(n)$ in the sense that it is also a low-pass filter with the same bandwidth denoted by $\bar{\omega}$. The frequency response of $a_{p,k}^{i,j}$ can be interpreted as the k-th frequency band of $a^{i,j}(n)$ multiplied by the frequency response of $\zeta_{p,k} = \zeta_k(n)|_{n=pD}$. One can see $\zeta_{p,k}$ is obtained by downsampling $\zeta_k(n)$ by the decimation factor D. The downsampling operation folds the frequency response with the period of $2\pi/D$. To avoid the frequency aliasing, the period should not be smaller than the bandwidth $\bar{\omega}$. For example, in this work, we use the Hamming window, the width of the main lobe is considered as the bandwidth, i.e. $\bar{\omega} = 8\pi/N$. Consequently, we set $D \leq N/4$. When $D < N/4$, the frequency response of $\zeta_{p,k}$ involves some side lobes, which have a magnitude close to zero. When $D = N/4$, only the main lobe is involved, and because the magnitude is dramatically decreasing from the center of the main lobe to its margin, the frequency region close to the margin of the main lobe has magnitude close to zero. This phenomenon that the frequency response of $\zeta_{p,k}$ and thus of $a_{p,k}^{i,j}$ are not fully excited is common to all microphones, which is problematic for solving (7). Fortunately, it is trivially known that the common zeros are introduced by the frequency response of $\zeta_{p,k}$. To make (7) solvable, we propose to determine the desired target $\mathbf{d}$ to have the same frequency response as $\zeta_{p,k}$, instead of the impulse function that has a full-band frequency response. To this end, the target $\mathbf{d}$ is designed as

$$\mathbf{d} = [0, \dots, 0, \boldsymbol{\zeta}^\top, 0, \dots, 0]^\top \in \mathbb{C}^{(L_a+L_h-1)\times 1}, \tag{9}$$

where $\boldsymbol{\zeta}$ denotes the vector form of $\zeta_{p,k}$. The zeros before $\boldsymbol{\zeta}$ introduce a modeling delay. As shown in [6], this delay is important for making the inverse filtering robust to perturbations of the CTF.

The solution of (7) gives an exact recovery of the j_d-th source plus the output noise as shown in (5). Following the proposition in [6], both the output noise and the influence of the CTF perturbations can be suppressed by reducing the energy of $\mathbf{h}$. This leads to the following optimization problem

$$\min_{\mathbf{h}} \parallel \mathbf{A}\mathbf{h} - \mathbf{g} \parallel^2 + \delta \phi_a^{j_d} \parallel \mathbf{h} \parallel^2, \tag{10}$$

where $\phi_a^{j_d} = \sum_{i=1}^{I} \sum_{p=0}^{L_a-1} |a_p^{i,j_d}|^2$ is the CTF energy for the desired source (summed over channels and frames), used as a normalization term, and δ is the regularization factor. Indeed, the power of the inverse filter $\mathbf{h}$ is at the level of $1/\phi_a^{j_d}$, thus $\parallel \mathbf{h} \parallel^2$ is somehow normalized by $\phi_a^{j_d}$. As a result, the choice of δ, which controls the trade-off between the two terms in (10), is unrelated to the power level of the CTF filters. This property is especially useful and necessary for the present frequency-wise algorithm where all the frequencies can share the same regularization factor δ, although the CTFs level may significantly vary along the frequencies.

The solution of (10), i.e. the CTF-domain inverse filter, is

$$\hat{\mathbf{h}}^{\text{mint}} = (\mathbf{A}^H\mathbf{A} + \delta\phi_a^{j_d}\mathbf{I})^{-1}\mathbf{A}^H\mathbf{g}, \tag{11}$$

where $\mathbf{I}$ is the IL_h-dimensional identity matrix.

4. EXPERIMENTAL EVALUATION

For experimental evaluation, we used the multi-channel impulse response data of [17], recorded using a 8-channel linear microphone array in a room of size 6 m × 6 m × 2.4 m. In the reported experiments, we used 6 microphones and RIRs with $T_{60} = 0.61$ s. The RIRs are truncated to correspond to T_{30}, and has a length of 5600 samples. The speech signals from the TIMIT dataset [18] are taken as the source signals, with the duration of about 3 s. A TIMIT speech is convolved with a RIR as the image of one source. Multiple (2, 3 or 4 in this experiment) image sources are summed as a mixture. For one mixture, the source direction and the

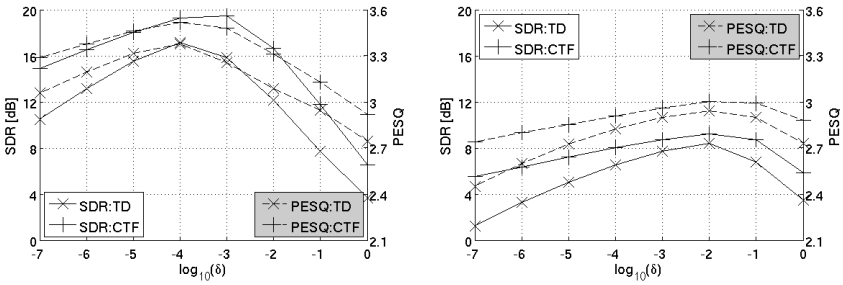

Fig. 1: Performance measures for 2-source as a function of δ. **left** NPM = -33 dB, **right** NPM = -15 dB. The SDR and PESQ of the input signals are -4.0 dB and 2.2, respectively. In legends, 'TD' denotes the time domain method.

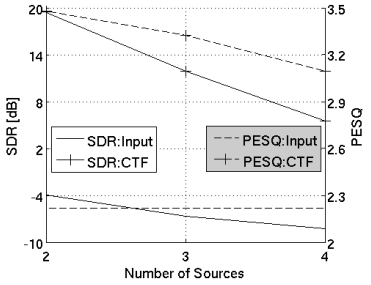

Fig. 2: Performance measures as a function of number of sources. NPM = -33 dB. $\delta = 10^{-3}$.

microphone-to-source distance of each source are randomly selected from $-90°$:$15°$:$90°$ and $\{1, 2\}$ m, respectively. To evaluate the robustness of the methods to the perturbations of the RIRs/CTFs, a proportional random Gaussian noise is added to the original filters $a^{i,j}(n)$ in the time domain to generate the perturbed filters denoted as $\tilde{a}^{i,j}(n)$. The noise level is denoted as the normalized projection misalignment (NPM) [19] in decibels (dB), i.e.

$$\text{NPM} = 10\log_{10}\frac{\sum_n (a^{i,j}(n) - \tilde{a}^{i,j}(n))^2}{\sum_n a^{i,j}(n)^2}.$$

Two NPM conditions, i.e. -33 dB and -15 dB are tested. The sampling rate is 16 kHz. The STFT uses the Hamming window, with the window length $N = 1,024$ (64 ms) and frame step $D = N/4 = 256$. The CTF length L_a is 29. The length of the inverse filter is set to $L_h = \lceil \frac{J(L_a-1)}{I-J} \rceil$, then $\mathbf{A}$ is square. The optimal setting of the modeling delay in $\boldsymbol{d}$ is related to the length of the inverse filters, i.e. L_h, and thus related to the number of sources. It is set to 4, 10, 18 taps for the cases of 2, 3 and 4 sources, respectively.

For each acoustic condition, 20 runs are performed, and the averaged perfomance measures are computed. The signal-to-distortion ratio (SDR) [20] is used to evaluate the overall quality of the outputs, especially the source separation performance. The perceptual evaluation of speech quality (PESQ) [21] is used to evaluate the quality of each individual output, especially the dereverberation performance. Note that the SDR of the input signals are computed for the mixture signals, the PESQ of the input signals is computed for each image source signal.

The time domain MINT [6] is taken as the baseline method, which is also set to recover the direct-path source signal with an energy regularization. In this experiment, we extend this method to the multisource case and set the parameters following the principles of the proposed method. For $J = 2$, the length of inverse filter and the modeling delay are set to 2800 and 1024, respectively. For $J = 3$ and 4, this method is not tested due to the requirement of large computation and memory resources.

Fig. 1 depicts the performance measures for 2-source mixtures as a function of the regularization factor δ. We can see that the performance significantly vary as a function of δ, and the best performance is achieved with different δ values for the two tested NPMs. This indicates that the regularization factor must be carefully tuned to achieve a good trade-off between the accuracy of the desired source recovery and the amplification of the filter perturbations. See [4] for further discussion on the optimal setting of δ. The proposed method achieves better SDR and PESQ scores than the time-domain MINT, despite of the fact that the CTF-based filtering is an approximation of the time-domain filtering. This is mainly due to much shorter filters in the STFT/CTF domain, and thus less sensitivity to filter misalignment. Moreover, shorter filters lead to a much smaller computation cost. In the present study, the computation time per mixture is 6 s for the proposed method, while it is 142 s for the time-domain MINT (both implemented in MATLAB). Overall, the proposed method achieves good performance measures for source separation and dereverberation: the SDR and PESQ (with the optimal δ) are 19.4 dB and 3.5 for NPM=-33 dB, 9.2 dB and 3.0 for NPM=-15 dB.

Fig. 2 shows the performance measures for various numbers of sources. It is seen that both SDR and PESQ scores decrease with the increase of the number of sources, however, a considerable performance improvement is achieved over input SDR and PESQ scores, even for the 4-source case. PESQ has a smaller degradation rate than SDR, which means that the perceptual quality of the desired source is relatively maintained in the presence of more interfering sources.

5. CONCLUSION

In this paper, a multisource MINT was proposed in the STFT domain, based on the CTF model. Experiments show that the proposed method is more efficient than the time-domain MINT in terms of both computation and performance. Overall, the proposed method is efficient for joint source separation and dereverberation.

6. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 2, pp. 145–152, 1988.

[2] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping-a study," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, vol. 5, pp. V101–V104, 2006.

[3] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, 2010.

[4] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.

[5] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, 2016.

[6] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.

[7] H. Yamada, H. Wang, and F. Itakura, "Recovering of broadband reverberant speech signal by sub-band MINT method," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 969–972, 1991.

[8] H. Wang and F. Itakura, "Realization of acoustic inverse filtering through multi-microphone sub-band processing," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 75, no. 11, pp. 1474–1483, 1992.

[9] S. Weiss, G. W. Rice, and R. W. Stewart, "Multichannel equalization in subbands," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 203–206, 1999.

[10] N. D. Gaubitch and P. A. Naylor, "Equalization of multichannel acoustic systems in oversampled subbands," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1061–1070, 2009.

[11] F. Lim and P. A. Naylor, "Robust speech dereverberation using subband multichannel least squares with variable relaxation," in *European Signal Processing Conference (EUSIPCO)*, 2013.

[12] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[13] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, 1992.

[14] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[15] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017.

[16] X. Li, L. Girin, and R. Horaud, "An em algorithm for audio source separation based on the convolutive transfer function," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.

[17] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *International Workshop on Acoustic Signal Enhancement*, pp. 313–317, 2014.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[19] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal processing letters*, vol. 5, no. 7, pp. 174–176, 1998.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.