NON-ITERATIVE IMPULSE RESPONSE SHORTENING METHOD FOR SYSTEM LATENCY REDUCTION

Jiawen Chua and W. Bastiaan Kleijn

School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

In this paper we present a non-iterative impulse response shortening method aiming to reduce the latency of a system. Our method exploits that smoothing the frequency-domain response generally leads to a shorter time-domain response. The method is simple to implement and has a computational complexity that is significantly lower than that of competing methods. Yet it achieves good performance. It can be used for applications involving system identification such as blind source separation (BSS), cross-talk cancellation and channel equalization. Our experimental results confirm the effectiveness of the method, demonstrating the benefit of the approach in the BSS and cross-talk cancelling applications.

Index Terms— Impulse response shortening, low latency, blind source separation, cross-talk cancellation

1. INTRODUCTION

In recent years, impulse response shortening methods have been studied widely for different applications. These include, but are not limited to, blind source separation (BSS) [1–3], speech dereverberation [4], channel equalization [5] and cross-talk cancellation [6, 7]. It has been used to minimize the artifacts resulting from circular convolution [1, 2] associated with the use of the fast Fourier transform and to compensate for the room reverberation with minimal latency [5–7]. We distinguish latency due to estimation computational time and due to algorithmic delay.

The algorithmic delay is of major importance for real-time applications involving system identification. In audio-processing applications, the channels contain memory, which results in a convolutive mixing process. To facilitate the estimation, the problem is usually transformed into the time-frequency (TF) domain using a short-time (ST) window. The channel in the frequency domain can then be independently estimated in every frequency bin. However, to account for the room reverberation, a long window is often needed. This prolongs the time taken for the signal acquisition, resulting in a system with a large algorithmic delay.

BSS is an example application where the latency problem occurs. The goal of blind source separation (BSS) is to extract the original sources from the observed mixtures with neither the prior knowledge of the mixing process nor the sources. Many TF-domain BSS approaches have been proposed. An overview of approaches can be found in [8, 9]. The traditional methods, including the well-known independent component analysis (ICA) approach [10] and the independent vector analysis (IVA) method [11], utilize the statistical properties to perform separation by assuming that the original sources are independent to each other. Later approaches employ the sparsity of the signals in the TF domain to estimate mixing matrices. These approaches include the degenerate unmixing estimation technique (DUET) [12], TIme-Frequency Ratio Of Mixtures (TIFROM) [13] and clustering algorithms [14]. Recently, the non-negative matrix factorization (NMF) approach [15] has also been introduced to solve the TF-domain BSS problem. Importantly, all these approaches assume that the ST window length is at least twice the room impulse response (RIR), which results in a high-resolution frequency domain (HRFD).

In practice, BSS approaches operating in the HRFD generally cannot be applied directly to real-time applications due to the delay between inputs and outputs. To resolve this problem, we [16] proposed a crossband filtering approach based on [17] to compute a low-resolution frequency-domain (LRFD) representation of the mixing filters using the HRFD mixing matrices. It reduces the time-lag of the system once the calculation is completed but the computational effort to estimate the crossband filters is high compared to that of the basic HRFD approach. Hence, the crossband filtering approach is only useful for stationary scenarios, where the source locations are fixed in time and the estimates can be updated infrequently.

The latency problem also occurs in cross-talk cancellation application. Cross-talk cancellation aims to deliver multiple signals to multiple listeners independently and simultaneously. Differently to BSS, the information of the RIRs is usually complete during the designation of the pre-filters. The pre-filters are used to process the signals before propagating to the listeners. The time-domain approaches [6,7] have been proposed to address this problem but suffer from high computational cost when the length of the RIRs increases.

Impulse response shortening can be applied to both the BSS and cross-talk cancellation applications to reduce the algorithmic latency. However, all the aforementioned approaches except [1] involve $\mathcal{L}_1$-norm or $\mathcal{L}_\infty$-norm minimization, which can lead to a slow convergence rate, prolonging the processing time. This again results in a system with long latency and it is impractical to real-time applications. Although [1] facilitates the finding of the optimal solution by using the least-square method, it is difficult to select the correct parameters and involves inversions of large matrices.

In this paper, we propose a non-iterative method to perform impulse response shortening to reduce the system latency. It reduces both the computational burden and shortens the time for signal acquisition. Our approach is based on the fact that the spectrum of a signal becomes smoother when the signal is zero-padded in the time-domain. Instead of computing the scaling factors by finding the sparsest representation of the RIRs in the time domain, we search for the complex scaling factors in the frequency domain that results in the smoothest spectrum. We demonstrate the advantage of the method in the BSS and cross-talk cancelling applications.

The paper is organized as follows. Section 2 introduces system identification model and our proposed method for impulse response shortening. The implementations of the proposed approach in BSS and cross-talk cancellation applications are discussed in Section 3 and Section 4, respectively. Section 5 presents the simulation results of our method. Conclusions are drawn in Section 6.

2. IMPULSE RESPONSE SHORTENING

In this section, we first review the model of system identification and discuss the approach for impulse response shortening of [1]. Next, we introduce our proposed method.

2.1. System identification model

We denote by H_m the frequency response of the RIR between the microphone m and a source, which is labelled as s. When the length of the ST window L_w is at least twice the length of the RIR L_h and it is generally considered that the linear system, the observation mixtures in the HRFD in a noiseless scenario can be approximated as [18]:

$$\mathbf{x}(p,k) \approx H(k)s(p,k), \tag{1}$$

where $\mathbf{x}(p,k)$ and $s(p,k)$ denote the vectors of the observations and the source, respectively, at time-frame index p and frequency bin k. $H(k)$ is a vector containing the frequency responses between all the microphones and the source, e.g., $\begin{bmatrix} H_1(k) & \cdots & H_m(k) \end{bmatrix}^{\mathrm{T}}$.

To shorten all the estimated RIRs, denoted as $\hat{H}$, a complex scaling factor $c(k)$ can be applied to $\hat{H}$ in each frequency bin to obtain a time-domain response. The complex scaling factor introduces a time-shift to each frequency component signal. When the factors are chosen suitably, the frequency component signals can be aligned such that a short time response is obtained. This causes a filtering effect to the recovered signal [19]. The filtering effect is generally not a significant issue in most applications. This is particularly true for BSS, where the scaling ambiguity is usually present in the frequency-domain approaches.

The approach of [1, 2] now uses a short-time Fourier transform (STFT) and the time-domain response can be expressed as $V_m\mathbf{c}$, where

$$V_m = \sum_l E_l \mathcal{F}^{-1} \mathrm{diag}(\mathcal{F} D_l \delta) \hat{H}_m, \tag{2}$$

and $\mathbf{c} = \begin{bmatrix} c(0), \cdots, c(L_w-1) \end{bmatrix}^{\mathrm{T}}$, is a vector of scaling factors. $\mathcal{F}$ is a discrete Fourier transform matrix and $\mathrm{diag}(\cdot)$ denotes an operator that converts a vector to a diagonal matrix. D_l, which is a diagonal matrix, defines the analysis window that is shifted to the l position. E_l is a shifting matrix that ensures the overlapping block are merged correctly. Note that the window length in D_l is shorter than L_w.

To perform impulse shortening, the authors [1] approximate the time-domain response to a pulse-like response, such as a delta function. The delta function is the shortest response as it contains only a single one and zeros otherwise. An optimal set of the complex scaling factor $\mathbf{c}$ can be obtained by minimizing

$$\left\| \mathbf{d} - \mathbf{V}\mathbf{c} \right\|_2, \tag{3}$$

where, $\left\| \cdot \right\|_2$ represents $\mathcal{L}_2$-norm. $\mathbf{V} = \begin{bmatrix} V_m^{\mathrm{T}} & \cdots & V_M^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ and $\mathbf{d} = \begin{bmatrix} d_1^{\mathrm{T}} & \cdots & d_M^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$, where d_m denotes the delta function that contains a single one at the peak position of $\hat{h}_m = V_m\mathbf{1}$. A trivial solution of (3) will be $\mathbf{c} = \mathbf{V}^{\dagger}\mathbf{d}$, where $\{\cdot\}^{\dagger}$ indicates a Moore-Penrose pseudoinverse.

In practice, (3) is not a good criterion to be minimized. It requires the estimated RIR $\hat{h}_m$ to be sparse, i.e., the amplitude of the main peak should be relatively large compared to the other peaks. This is not the case in some applications, e.g. BSS, as the estimated mixing matrices suffer from the so-called scaling ambiguity across the spectrum. More details will be discussed in Section 3. In addition, it needs to perform an inverse operation of large matrices, which increases the computational burden.

2.2. Proposed method

In this subsection, we propose a non-iterative method to perform impulse response shortening. Similarly to [1, 2], we introduce a new set of complex scaling factors to shorten the estimated RIRs.

The motivation for our method is based on the fact that zero-padding a signal in the time domain leads to a smoother frequency spectrum. We hypothesize that this implies that smoothing the frequency spectrum leads to a short response. Instead of finding the scaling factors by making the estimated RIRs sparse in the time domain as done in [1, 2], we search for a new set of complex scaling factors in the frequency domain that lead to the smoothest spectrum.

A smooth spectrum can be obtained by altering the coefficients of the frequency response $\hat{H}_k$, such that they are maximally similar to each other in adjacent bins. This can be done by approximating the coefficients to the ones in the previous frequency bin:

$$\left\| H(k-1) - c(k)H(k) \right\|_2, \quad k = [1, L_w - 1] \tag{4}$$

Hence, the frequency response $\tilde{H}_k$ that varies smoothly in frequency can be computed as:

$$\tilde{H}(k) = \begin{cases} \frac{\hat{H}(k)}{\left\| \hat{H}(k) \right\|} & k = 0 \\ \frac{c(k)\hat{H}(k)}{\left\| c(k)\hat{H}(k) \right\|} & k = [1, L_w - 1] \end{cases}, \tag{5}$$

$$c(k) = \frac{\hat{H}(k)^{\mathrm{H}} \tilde{H}(k-1)}{\hat{H}(k)^{\mathrm{H}} \hat{H}(k)}, \tag{6}$$

where $\{\cdot\}^{\mathrm{H}}$ denotes a Hermitian transpose.

Our proposed approach differs from [1, 2] as the desired time-domain responses are not required. It does not involve matrix inversion. Unlike [6, 7], our method is non-iterative. Hence, the computational efficiency is higher, and is simpler to implement. We show that our proposed method can be implemented to design both the post-filters as shown in Section 3 for BSS applications and the pre-filters as presented in Section 4 for cross-talk cancellation implementations.

3. BLIND SOURCE SEPARATION APPLICATION

This section first reviews the formulation of the BSS problem in the TF domain. Next, we apply our proposed approach for impulse response shortening. This facilitates the computation of the representations of the LRFD mixing filters. Then, we estimate the demixing operator in the LRFD based on the shortened mixing filters.

3.1. Problem formulation of BSS

In this subsection, we first briefly provide the necessary background for BSS. We neglect the effect of noise in the derivation and consider the overdetermined scenario, where the number of microphones M is larger than the number of original sources N, i.e. $M > N$.

The observation mixtures in the HRFD can be written as:

$$\mathbf{x}(p,k) \approx A(k)\mathbf{s}(p,k), \tag{7}$$

where $\mathbf{x}(p,k)$ and $\mathbf{s}(p,k)$ now denote the vectors of the observation mixtures and the original sources at time-frame index p and frequency bin k, respectively, while $A(k)$ represents the mixing matrix at frequency bin k. The mixing matrices $A(k)$ (or, alternatively, the demixing matrices), can be estimated using the aforementioned BSS algorithms [10–15].

To minimize the algorithmic latency, the BSS problem can be solved by using a crossband filtering approach [16]. The approach designs the demixing operators based on the mixing filters in the LRFD, which are obtained from the HRFD mixing matrices. Hence, the methods that aim to estimate $A(k)$ [12–14] are preferred.

The estimation delay is large as the computational effort is extremely high if the estimated mixing matrix, denoted as $\hat{A}$, is directly used in [16]. Due to the scaling ambiguity, a random complex scaling factor is introduced into each column of the HRFD mixing matrix in every frequency bin. This causes a random time-shift in each frequency band signal and leads to long RIRs.

The computational efficiency in [16] can be improved by shortening and truncating the mixing filters beforehand. In this case, the effect of the crossband filters will be minimal and can be neglected. This significantly improves the calculation speed. The details will be discussed in the next subsection.

3.2. Estimating the demixing operator

We follow the method described in [16] to estimate the demixing operators in the HRFD, which is based on the LRFD mixing filters. To perform separation, we need to compute the representation of the shortened mixing filters in the LRFD. The shortened mixing filters can be obtained by using the proposed method in Section 2.2. It is done by repeating (5) and (6) for each column of $A(k)$ for every frequency bin, e.g. $\hat{H}(k)$ is replaced by $\hat{A}_{\cdot n}(k)$ and $\tilde{H}(k)$ is replaced by $\tilde{A}_{\cdot n}(k)$, where $Z_{\cdot n}$ indicates the n^{th} column of the matrix Z.

To facilitate the computation in [17], the length of the truncated RIRs is desired to be the length of the ST window in the LRFD. This is to diminish the effect of the crossband filters. The truncation can be done by applying a rectangular window to capture the segments containing the highest $\mathcal{L}_1$-norm, so that the maximal information is retained. Without loss of generality, we rotate the RIRs, such that the segments are located in the middle. The truncated RIRs can be represented as:

$$\bar{\mathcal{A}}_n = D_n^* \tilde{\mathcal{A}}_n, \tag{8}$$

$$D_n^* = \underset{D_n \in \mathcal{D}}{\arg\max} \quad \mathbf{1}^{\mathrm{T}} \left| D_n \tilde{\mathcal{A}}_n \right| \mathbf{1}, \tag{9}$$

where $\hat{\mathcal{A}}_n = \begin{bmatrix} \hat{a}_{\cdot n}(0) & \cdots & \hat{a}_{\cdot n}(L_w - 1) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{L_w \times M}$ contains the shortened RIRs between the n^{th} source and all the M microphones in the time domain. $\mathcal{D}$ is a set of zero matrices with an identity matrix located at the i^{th} column index:

$$\mathcal{D} = \left\{ \begin{bmatrix} \mathbf{0}^{L_s \times i} & \mathbf{I}^{L_s \times L_s} & \mathbf{0}^{L_s \times (L_w - L_s - i)} \end{bmatrix} \;\middle|\; i \in [0, L_w - L_s] \right\}, \tag{10}$$

where L_s is the length of the ST window in the LRFD.

The truncated RIRs $\bar{a}_{mn}$ can be found by computing (8) and (9) repeatedly for $n = [1, N]$. Then, we calculate the LRFD representation of the band-to-band mixing filters [17] and design the demixing operators in the LRFD using the method described in Section 4 [16].

4. CROSS-TALK CANCELLATION APPLICATION

This section first briefly reviews the problem definition of the cross-talk cancellation. Then, we implement the proposed response shortening approach to design pre-filters for cross-talk cancellation.

The objective of the cross-talk cancellation is to deliver the signals from Q loudspeakers to R listeners independently and simultaneously. In general, $Q \geq R$. This can be done by designing $R \times Q$ pre-filters to compensate for the $Q \times R$ RIRs between the loudspeakers and the listeners.

Let us define g_{rq} as the time-domain pre-filter that compensates for the signal, which propagates from the q^{th} loudspeaker to the r^{th} listener. Let us denote by b_{qr} the RIR between the loudspeaker q and the listener r in the time domain. A time-domain approach to finding the pre-filters can be found in [6, 7]. Both methods involve norm minimization and are implemented iteratively, which can lead to a reduced convergence rate and prolong the estimation delay.

We simplify the problem by transforming the problem into the frequency domain. Hence, the pre-filter can be estimated in every frequency bin in the HRFD independently:

$$G(k)B(k) = \mathbf{I}, \tag{11}$$

where $G(k) \in \mathbb{C}^{R \times Q}$ and $B(k) \in \mathbb{C}^{Q \times R}$ represent the frequency responses of the pre-filters and the RIRs in frequency bin k, respectively.

To shorten the pre-filters, we apply the approach proposed in Section 2. Although the scaling ambiguity is not present in the cross-talk-cancellation application (the RIRs are generally fully known), the performance is generally not significantly affected. The shortened pre-filters can be obtained by substituting $G_{r\cdot}^{\mathrm{T}}(k) = \hat{H}(k)$ and $\tilde{G}_{r\cdot}^{\mathrm{T}}(k) = \tilde{H}(k)$ in (5) and (6) for each row of $G(k)$ for every frequency bin.

5. RESULTS

In this section, we discuss the experimental results for impulse response shortening for both the BSS and cross-talk cancellation applications. In both cases, we first provide the setup and then the simulation results. We note that the main focus of this paper is to reduce the system latency. Hence, we focus on the computational effort and the separation performance in the performance evaluation.

5.1. BSS experimental setup

Three 10 second speech signals sampled at 16 kHz, which were obtained from Stereo Audio Source Separation Evaluation Campaign (SASSEC) [20], were used. We altered the activity period, such that, for each source, certain periods existed where only one source was active. This is not necessary but guarantees that the mixing matrices can be correctly estimated using the sparsity-based method [12–14]. All the separations were conducted offline and computed by Matlab R2015b on a PC having an Intel(R) Core(TM) i5-5200 CPU@2.20 GHz processor with 8GB random-access memory.

In the simulation, 24 microphones were used. The observations were obtained by convolving the speech signals with the simulated room impulse responses (RIRs). 24×3 RIRs with 1024 taps were computed using the image-source method [21], where the microphones and sources were randomly placed in a room with a size of $3\ \mathrm{m} \times 3\ \mathrm{m} \times 3\ \mathrm{m}$. The reverberation time was 0.2 s.

Square-root of Hanning windows with 2048 taps (128 ms) and 512 taps (32 ms) were used in the high-resolution frequency domain (HRFD) and in the low-resolution frequency domain (LRFD), respectively. The windows were 50% overlapped. The HRFD mixing matrices were estimated using both the ICA [10] and Modified-TIFROM [13, 14] approaches. The explanation of the Modified-TIFROM method can be found in [16]. The permutation ambiguity was resolved using oracle information, so that the separation performance was not affected by the permutation issue.

We examined five different approaches for each estimation method. The stand-alone strategy identified the demixing matrices in the LRFD directly while the crossband approach [16] utilized

Table 1: Performance comparison between various approaches for BSS using simulated data.

Metrics	ICA					Modified-TIFROM				
	Stand-alone	Crossband [16]	Minimal [3]	Mazur [1]	Proposed	Stand-alone	Crossband [16]	Minimal [3]	Mazur [1]	Proposed
SIR_1 (dB)	9.70	16.47	13.62	8.24	**18.32**	5.31	16.24	14.62	17.77	**18.72**
SIR_2 (dB)	10.90	12.42	15.64	13.18	**16.20**	20.14	22.31	22.87	19.25	**25.45**
SIR_3 (dB)	10.69	**19.13**	13.66	15.74	15.60	12.82	18.44	16.04	21.07	**21.26**
SIR_{avg} (dB)	10.43	16.01	14.31	12.38	**16.71**	12.76	18.99	17.84	19.36	**21.81**
Time (s)	**5.39**	66.71	20.52	158.35	20.37	**5.01**	63.45	18.13	152.72	18.24

the HRFD mixing matrices to compute the LRFD mixing filters and designed the LRFD demixing operators. In addition to our proposed approach, we tested two different impulse response shortening methods for comparison. The minimal approach shortens the estimated RIRs by resolving the scaling ambiguity based on the minimal distortion principle [3]. The method proposed by Mazur et. al. [1] is described in Section 2. It obtains an optimal set of complex scaling factors by approximating the time-domain shortened filters to a desired pulse-like response. After shortening, the estimated HRFD mixing filters were truncated as suggested in Section 3.2 and the demixing operators were designed based on the LRFD representation of the truncated mixing filters.

5.2. BSS simulation results

The BSS_EVAL toolbox [22] was used to compute the signal-to-interference ratio (SIR) between the separated source and the original source to indicate the source separation performance. A higher score indicates better performance. We also compare the computation time of each method. The results of all the approaches are tabulated in 1. SIR_l represents the SIR of the l^{th} while SIR_{avg} indicates the average value of the SIR values in each method.

The results show that the proposed approach achieves the highest average SIR value using both the ICA approach and the M-TIFROM method. In terms of the computation time for the estimation of the response, the proposed approach was slower than the stand-alone method but is significantly faster than the state-of-the-art procedures.

5.3. Cross-talk cancellation experimental setup

In the simulation, four loudspeakers and two microphones, acting as listeners, were randomly located in a room with a size of $3\,\text{m} \times 3\,\text{m} \times 3\,\text{m}$. The room impulse responses (RIRs) with 1024 taps were generated using the image source method [21], where the reverberation time was 0.2 s. The sampling rate was 16 kHz.

We compare the proposed approach with [6]. The $\mathcal{L}_2$-norm was chosen as a criterion to be minimized to facilitate the computation in [6].

5.4. Cross-talk cancellation simulation results

The performance of the cross-talk cancellation is measured using a direct signal to cross-talk ratio (DSCR) [23], which is the ratio of the maximum direct response to the maximum cross-talk. High DSCR value indicates good cross-talk cancellation performance. Fig. 1 shows the cross-talk cancellation performance of our approach. The DSCR values between the direct response and the cross-talk for the first signal and for the second signal were 17.81 dB and 17.59 dB, respectively. The computational time was 0.10 s. For the approach

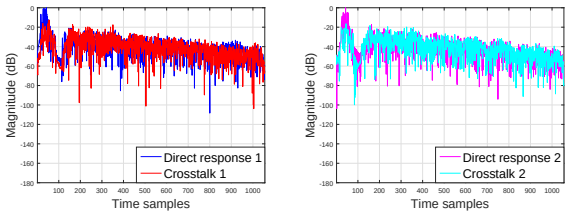

Fig. 1: Cross-talk cancellation performance of the proposed approach, where $Q = 4$ and the pre-filter length was 128.

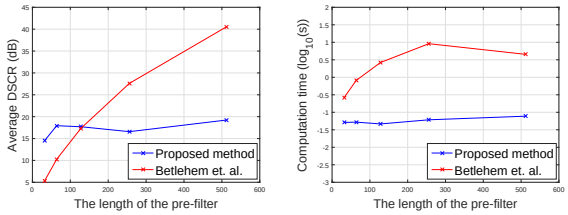

Fig. 2: Performance comparison with different pre-filter length.

in [6] with the same setup, the DSCR values were 17.67 dB and 16.98 dB while the computational time was 3.26 s. Both methods obtained similar DSCR values but our proposed method was significantly faster than the method of [6].

In Fig. 2, we compare the performance of the two approaches against the length of the pre-filters. The results show that [6] surpasses our proposed approach when the filter length is longer. This is a consequence of the fact that the proposed method designs the pre-filters in the frequency domain, resulting in the distortion associated with circular convolution. Long filters make this distortion more severe. However, our method is more efficient than [6] independently of the length of the pre-filters.

6. CONCLUSION

In this paper, we presented a non-iterative impulse response shortening method. It aims to reduce the system latency, including both the signal acquisition time and the processing time. The method provides a simple and practical solution for real-time applications involving system identification. The simulation results show that the approach can be used to design both post-filters for the BSS application and pre-filters for the cross-talk cancellation application. Future work may include an investigation of the perception of the filtering effect.

7. REFERENCES

[1] R. Mazur and A. Mertins, "Using the scaling ambiguity for filter shortening in convolutive blind source separation," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 1709–1712.

[2] R. Mazur and A. Mertins, "A method for filter shaping in convolutive blind source separation," in *Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings*, Berlin, Heidelberg, 2009, pp. 282–289, Springer Berlin Heidelberg.

[3] K. Matsuoka, "Minimal distortion principle for blind source separation," in *SICE 2002. Proceedings of the 41st SICE Annual Conference*. IEEE, 2002, vol. 4, pp. 2138–2143.

[4] W. Zhang, A. W. H. Khong, and P. A. Naylor, "Acoustic system equalization using channel shortening techniques for speech dereverberation," in *2009 17th European Signal Processing Conference*, Aug 2009, pp. 1427–1431.

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

[6] T. Betlehem, P. D. Teal, and Y. Hioka, "Efficient crosstalk canceler design with impulse response shortening filters," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 393–396.

[7] L. Krishnan, P. D. Teal, and T. Betlehem, "A robust sparse approach to acoustic impulse response shaping," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 738–742.

[8] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 7, pp. 1640–1655, 2005.

[9] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Multichannel Speech Processing Handbook*, pp. 1065–1084. New York, NY, USA: Springer, 2007.

[10] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[11] T. Kim, T. Eltoft, and T. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Independent Component Analysis and Blind Signal Separation*, pp. 165–172. Berlin Heidelberg: Springer, 2006.

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

[13] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403, 2005.

[14] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time-frequency masking.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 1, pp. 101–116, 2010.

[15] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[16] J. Chua, G. Wang, and W. B. Kleijn, "Convolutive blind source separation with low latency," in *International Workshop on Acoustic Signal Enhancement (IWAENC' 16)*, September 2016.

[17] 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.

[18] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC '01)*. September 2001, Darmstadt, Germany.

[19] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 200–200, 2006.

[20] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Independent Component Analysis and Signal Separation*, pp. 552–559. Berlin Heidelberg: Springer, 2007.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[22] 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.

[23] J. O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1829–1842, 2012.