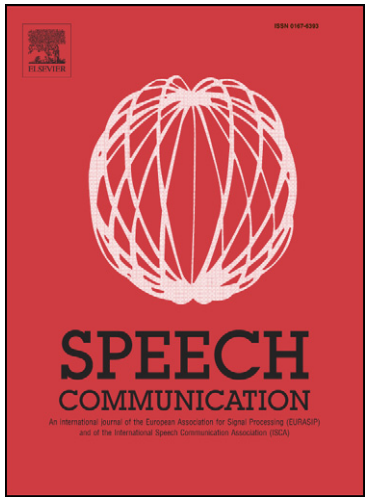
A Multistage Approach to Blind Separation of Convolutive Speech Mixtures

Tariqullah Jan, Wenwu Wang, Deliang Wang

Accepted Manuscript



A Multistage Approach to Blind Separation of Convolutive Speech Mixtures

Tariqullah Jan, Wenwu Wang, DeLiang Wang

PII: S0167-6393(11)00003-3
DOI: 10.1016/j.specom.2011.01.002
Reference: SPECOM 1960

To appear in: *Speech Communication*

Received Date: 10 August 2010
Revised Date: 23 December 2010
Accepted Date: 3 January 2011

Please cite this article as: Jan, T., Wang, W., Wang, D., A Multistage Approach to Blind Separation of Convolutive Speech Mixtures, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.01.002

A Multistage Approach to Blind Separation of Convolutive Speech Mixtures

Tariqullah Jan†, Wenwu Wang†, and DeLiang Wang‡

† *Centre for Vision, Speech and Signal Processing, University of Surrey, UK*
Email: {t.jan, w.wang}@surrey.ac.uk
‡ *Department of Computer Science and Engineering & Centre for Cognitive Science, The Ohio State University, Columbus, USA*
Email: dwang@cse.ohio-state.edu

.

Abstract

We propose a novel algorithm for the separation of convolutive speech mixtures using two-microphone recordings, based on the combination of independent component analysis (ICA) and ideal binary mask (IBM), together with a post-filtering process in the cepstral domain. The proposed algorithm consists of three steps. First, a constrained convolutive ICA algorithm is applied to separate the source signals from two-microphone recordings. In the second step, we estimate the IBM by comparing the energy of corresponding time-frequency (T-F) units from the separated sources obtained with the convolutive ICA algorithm. The last step is to reduce musical noise caused by T-F masking using cepstral smoothing. The performance of the proposed approach is evaluated using both reverberant mixtures generated using a simulated room model and real recordings in terms of signal to noise ratio measurement. The proposed algorithm offers considerably higher efficiency and improved speech quality while producing similar separation performance compared with a recent approach.

Key words: Independent component analysis (ICA), convolutive mixtures, ideal binary mask (IBM), estimated binary mask, cepstral smoothing, musical noise.

1 Introduction

1.1 Problem Description and Previous Work

The extraction of a target speech signal from a mixture of multiple signals is classically referred to as the cocktail party problem (Cherry, 1953). Although

it poses big challenges in many signal processing applications, human listeners with normal hearing are generally very skillful in separating the target speech from a complex auditory scene (Wang and Brown, 2006). Listeners with hearing loss suffer from insufficient speech intelligibility in noisy environments (Dillon, 2001). Simply amplifying the input is not sufficient to increase the intelligibility of the target speech as both the target and interfering signals are amplified. Despite being studied for decades, the cocktail party problem remains a scientific challenge that demands further research efforts (Wang and Brown, 2006). Computational modelling and algorithmic solutions to this problem are likely to have strong impact on several applications including hearing aids and cochlear implants, human-machine interaction and robust speech recognition in uncontrolled natural environments.

One promising technique to address this problem for convolutive mixtures [1] is the framework of blind source separation (BSS) where the mixing process is generally described as a linear convolutive model, and independent component analysis (ICA) (Hyvarinen et al., 2001; Lee, 1998) can then be applied to separate the convolutive mixtures either in the time domain (Cichocki and Amari, 2002; Douglas and Sun, 2002; Douglas et al., 2007), in the transform domain (Araki et al., 2003; Makino et al., 2005; Olsson and Hansen, 2006; Rahbar and Reilly, 2005; He et al., 2007; Reju et al., 2010; Aissa-El-Bey et al., 2007; Wang et al., 2005; Yoshioka et al., 2009; Han et al., 2009), or their hybrid (Lambert and Bell, 1997; Lee et al., 1997), assuming the source signals are statistically independent (Araki et al., 2003; Douglas et al., 2005; Makino et al., 2005; Mitianondis and Davies, 2002; Nickel and Iyer, 2006; Olsson and Hansen, 2006). The time-domain approaches attempt to extend instantaneous ICA methods for the convolutive case. Upon convergence, these algorithms can achieve good separation performance due to the accurate measurement of statistical independence between the segregated signals (Makino et al., 2005). However, the computational cost associated with the estimation of the filter coefficients for the convolution operation can be very demanding, especially when dealing with reverberant (or convolutive) mixtures using filters with long time delays (Amari et al., 1997; Buchner et al., 2004; Douglas and Sun, 2002; Matsuoka and Nakashima, 2001).

To reduce computational complexity, the frequency-domain approaches transform the time-domain convolutive model into a number of complex-valued instantaneous ICA problems, using the short-time Fourier transform (STFT) (Araki et al., 2003; Mukai et al., 2004; Parra and Spence, 2000; Sawada et al., 2003; Schobben and Sommen, 2002; Wang et al., 2005). Many well-established instantaneous ICA algorithms can then be applied at each frequency bin. Nev-

[1] In speech separation, the term "convolutive mixtures" refers to the signals received by microphones that are from multiple speakers in an environment with surface reflections from e.g. walls, ceilings and floors.

ertheless, an important issue associated with this approach is the so-called permutation problem (Sawada et al., 2004; Wang et al., 2004), i.e., the permutation of the source components at each frequency bin may not be consistent with each other. As a result, the estimated source signals in the time domain (using an inverse STFT transform) may still contain the interferences from the other sources due to the inconsistent permutations across the frequency bands. Different methods have been developed to solve the permutation problem, such as the filter length constraint approaches (Buchner et al., 2004; Parra and Spence, 2000), the source localization or beamforming approaches (Sawada et al., 2004; Soon et al., 1993), the method based on the physical behaviour of the acoustic environment (Nesta et al., 2008) or coherent source spectral estimation (Nesta et al., 2009), the approach for modeling frequency bins using the generalized Gaussian distribution (Mazur and Mertins, 2009) and the method based on the envelope correlation between the estimated source components at the frequency bins (Murata et al., 2001).

Hybrid time-frequency (T-F) methods tend to exploit the advantages of both time- and frequency-domain approaches, and consider the combination of the two types of methods. In particular, the coefficients of the FIR filter are typically updated in the frequency domain and the non-linear functions are adopted in the time domain for evaluating the degree of independence between the source signals (Back and Tosi, 1994; Lee et al., 1997). In this case, no permutation problem exists any more, as the independence of the source signals is evaluated in the time domain. Nevertheless, a limitation with the hybrid approaches is the increased computational load induced by the back and forth movement between the two domains at each iteration using the Discrete Fourier transform (DFT) and inverse DFT (Makino et al., 2005).

Although the convolutive BSS problem, i.e. separating unknown sources from their convolutive mixtures, has been studied extensively, the separation performance of developed algorithms is still limited, and leaves much room for further improvement. This is especially true when dealing with reverberant and noisy mixtures. For example, in the frequency-domain approaches, if the frame length for computing the STFT is long and the number of samples within each window is small, the independence assumption may not hold any more (Araki et al., 2003). On the other hand, a short size of the STFT frame may not be adequate to cover the room reverberation, especially for mixtures with long reverberations for which a long frame size is usually required for keeping the permutations consistent across the frequency bands (see e.g., (Back and Tosi, 1994; Lee et al., 1997)).

A recent technique proposed in computational auditory scene analysis (CASA), called ideal binary mask (IBM), has shown promising properties in suppressing interference and improving intelligibility of target speech. IBM is obtained by comparing the T-F representations of target speech and back-

ground interference, with 1 assigned to a T-F unit where the target energy is stronger than the interference energy and 0 otherwise (Wang, 2005). The target speech can then be obtained by applying the IBM to the T-F representation of the mixture, together with an inverse transform. The IBM technique was originally proposed as a computational goal or performance benchmark of a CASA system (Wang, 2005; Wang and Brown, 2006). Recent studies reveal that by suppressing the interference signals from the mixtures, the IBM technique can significantly improve the intelligibility of the target speech (Wang et al., 2009). This simple yet effective approach offers great potential for improving speech separation performance of ICA algorithms. Different from many ICA approaches with linear models (Madhu et al., 2008), signals estimated in the T-F plane have mostly non-overlapping supports for different speaker signals and thus one can use IBM to extract the target speech from their mixture signal. The IBM is obtained by assuming both the target speech and interfering signal are known *a priori*. However, in practice, only mixtures are available, and the IBM must be estimated from the mixtures, which is a major computational challenge. Several CASA methods have been developed for this purpose, see e.g., (Roman et al., 2003; Wang and Brown, 2006; Rodrigues and Yehia, 2009).

Recently Pedersen *et al.* (Pedersen et al., 2008) proposed to estimate the IBM from intermediate separation results that are obtained by applying an ICA algorithm to the mixtures. The limitation of the aforementioned CASA methods, i.e., having to estimate the IBM directly from the mixtures, is mitigated as the IBM can now be estimated from the coarsely separated source signals obtained by ICA algorithms. The estimated IBM can be further used to enhance the separation quality of the coarsely separated source signals. Such a combination was shown to achieve good separation performance. However, both the mixing model and separation algorithm considered in (Pedersen et al., 2008) are instantaneous, which in practice may not be sufficient for real recordings. Related work was proposed in (Sawada et al., 2006) where the target speech is extracted from the mixture using ICA and time-frequency masking. However the errors introducted in the estimation of the binary T-F mask have not been addressed. In this paper, we explore the combination of ICA and IBM techniques for the separation of convolutive speech mixtures by using a convolutive mixing model and a convolutive separation algorithm. To deal with the estimation errors of the binary mask, we employ a cepstrum based processing method.

1.2 Overview of Proposed Method

In our proposed algorithm, we first apply a constrained convolutive ICA algorithm (Wang et al., 2005) to the microphone recordings. As is common with

many other existing ICA algorithms, the separated target speech from this step still contains a considerable amount of interference from other sources. The performance steadily degrades with an increase of reverberation time. In order to reduce the interference within the target speech, we estimate the IBM by comparing the energy of the corresponding T-F units from the outputs of the convolutive ICA algorithm, and then apply the estimated IBM to the original mixtures to obtain the target speech and interfering sources. As will be confirmed in our experiments, this process considerably improves the separation performance by reducing the interference to a much lower level. However, a typical problem with the binary T-F masking is the introduction of errors in the estimation of the masks. The errors may result in some isolated T-F units, causing fluctuating musical noise (Araki et al., 2005; Madhu et al., 2008).

In this paper, we propose to reduce such noise by further processing the estimated IBM using cepstral smoothing (Madhu et al., 2008). More specifically, we transform the binary mask into the cepstral domain, and smooth the transformed mask over time frames using the overlap-and-add technique. In the cepstrum domain, it is easier to distinguish between the unwanted isolated random peaks and mask patterns resulting from the spectral structure of the segregated speech. Therefore, different levels of smoothing can be applied to the binary T-F mask in different frequency ranges. The smoothed mask, after being transformed back into the T-F plane, is then applied to the outputs of the previous step in order to reduce the musical noise.

Our proposed approach is essentially a multistage algorithm, as depicted by a block diagram in Figure 1 for two microphone mixtures. In the first stage, convolutive speech mixtures $x_1(n)$ and $x_2(n)$ are processed by the constrained convolutive ICA algorithm in (Wang et al., 2005), where n represents the discrete time index. The resultant estimated source signals of this stage are denoted as $y_1(n)$ and $y_2(n)$. In the second stage, the T-F representations of $y_1(n)$ and $y_2(n)$ are used to estimate the IBM, and the resultant masks are denoted by $M_1^f(k,m)$ and $M_2^f(k,m)$, where k represents the frequency index, and m is the time frame index. The final stage is to perform smoothing of the estimated IBM in the cepstral domain to reduce the musical noise. The smoothed version of the estimated IBM is denoted by $\overline{M}_1^f(k,m)$ and $\overline{M}_2^f(k,m)$, as shown in Figure 1. Finally, the smoothed masks (after being converted back to the spectral domain) are applied to the outputs of the previous step, followed by an inverse T-F transform to obtain the estimated source signals in the time domain. The details of each step are described in the following sections. A preliminary version of this work was presented in (Jan et al., 2009).

The remainder of the paper is organised as follows. The convolutive ICA approach and its utilization in the first stage of our proposed method is presented in Section 2. Section 3 describes in detail the second stage of the algorithm, i.e., how to estimate the IBM from the outputs of the convolutive ICA algo-

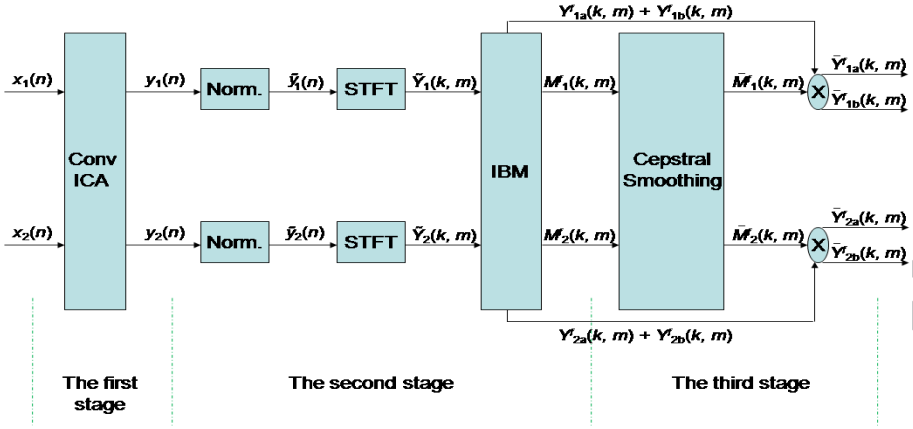


Fig. 1. Block diagram of the proposed multistage approach. In the first stage, a convolutive ICA algorithm (denoted as "Conv ICA") is applied to the mixture signals $x_j(n)$ $(j = 1, 2)$ to obtain the coarsely separated signals $y_i(n)$ $(i = 1, 2)$. In the second stage, $y_i(n)$ is first normalised (denoted as "Norm") to obtain $\tilde{y}_i(n)$, which is then transformed to $\tilde{Y}_i(k, m)$ using the STFT followed by the estimation of the binary masks $M_i^f(k, m)$. In the third stage, cepstral smoothing is applied to the estimated masks $M_i^f(k, m)$ and the smoothed masks $\overline{M}_i^f(k, m)$ are then used to enhance the separated speech signals obtained from the second stage.

rithm. Musical noise reduction using cepstral smoothing, i.e., the final stage of the proposed algorithm, is explained in Section 4. Section 5 thouroughly evaluates the proposed method and compares it with two recent methods (Pedersen et al., 2008) and (Wang et al., 2005). Further discussions about the results and some conclusions are given in Section 6.

2 BSS of Convolutive Mixtures in the Frequency Domain

In a cocktail party environment, N speech signals are recorded by M microphones, which can be described mathematically by a linear convolutive model

$$x_j(n) = \sum_{i=1}^{N} \sum_{p=1}^{P} h_{ji}(p) s_i(n - p + 1) \qquad (j = 1, ..., M) \tag{1}$$

where s_i and x_j are the source and mixture signals respectively, h_{ji} is a P-point room impulse response (Gaubitch, 1979) from source s_i to microphone x_j. The BSS problem for convolutive mixtures in the time domain is converted to multiple instantaneous problems in the frequency domain by applying the short time Fourier transform (STFT) to equation (1), see e.g. (Smaragdis, 1998; Parra and Spence, 2000; Sawada et al., 2004, 2007; Rahbar and Reilly, 2005; Araki et al., 2003; He et al., 2007; Reju et al., 2010; Aissa-El-Bey et al.,

2007; Wang et al., 2005; Yoshioka et al., 2009; Han et al., 2009), and using matrix notations, as follows

$$\mathbf{X}(k,m) = \mathbf{H}(k)\mathbf{S}(k,m) \tag{2}$$

where $\mathbf{X}(k,m) = [X_1(k,m), ..., X_M(k,m)]^T$ with its elements $X_j(k,m)$ being the T-F representations of the microphone signals $x_j(n)$, $\mathbf{S}(k,m) = [S_1(k,m), ..., S_N(k,m)]^T$ whose elements $S_i(k,m)$ are the T-F representations of the source signals $s_i(n)$, and $[\cdot]^T$ denotes vector transpose. The mixing matrix $\mathbf{H}(k)$ is assumed to be invertible and time invariant. In this study we consider a two-input two-output system, i.e., $N = M = 2$.

To find the sources, we can apply an unmixing filter $\mathbf{W}(k)$ to the mixtures, also shown in Figure 2

$$\mathbf{Y}(k,m) = \mathbf{W}(k)\mathbf{X}(k,m) \tag{3}$$

where $\mathbf{Y}(k,m) = [Y_1(k,m), Y_2(k,m)]^T$ represents the estimated source signals in the T-F domain and $\mathbf{W}(k)$ is denoted as $[[W_{11}(k), W_{12}(k)]^T, [W_{21}(k), W_{22}(k)]^T]^T$, which can be estimated based on the assumption of independence. Many algorithms have been developed for this purpose (Araki et al., 2004, 2003, 2007; Cichocki and Amari, 2002; Parra and Spence, 2000; Sawada et al., 2007). In this work we use a constrained convolutive ICA approach in (Wang et al., 2005) for the estimation of $\mathbf{W}(k)$. Applying an inverse STFT (ISTFT), $\mathbf{Y}(k,m)$ can be converted back to the time domain denoted as

$$\mathbf{y}(n) = \text{ISTFT}(\mathbf{Y}(k,m)) \tag{4}$$

where $\mathbf{y}(n) = [y_1(n), y_2(n)]^T$ denotes the estimated source signals in the time domain. This inverse transform is for the purpose of applying a scaling operation to the estimated sources, as explained in the next section. Similar to many existing ICA approaches, e.g., (Parra and Spence, 2000), however, the separation performance of (Wang et al., 2005), especially the quality of the separated speech, is still limited due to the existence of a certain amount of interference within the separated speech. The performance further degrades with an increase of the reverberation time (RT). Such degradation is caused partly by the tradeoff between the filter length used in the convolutive model and the frame length of the STFT within the frequency-domain algorithms. For a high reverberation condition, an unmixing filter with long time delays is usually preferred for covering sufficiently the late reflections. On the other hand, the frequency domain operation usually requires the frame length of the STFT to be significantly greater than the length of the unmixing filter, in order to keep the permutation ambiguities across the frequency bands to a minimum. The filter length constraint may be relaxed when other techniques, such as beamforming and source envelope correlations (Murata et al., 2001;

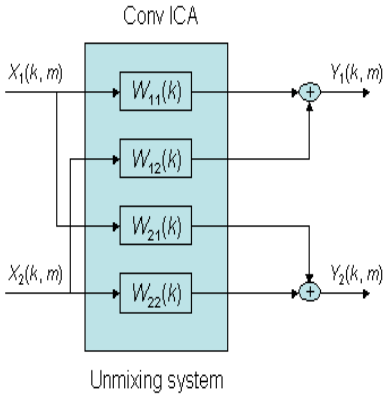


Fig. 2. Block diagram showing the first stage of the proposed approach. The mixture signals $x_j(n)$ $(j = 1, 2)$ are first transformed into the T-F domain using the discrete STFT. The resultant T-F representation $X_j(k, m)$, as the input to a frequency-domain BSS algorithm, is then used to estimate the unmixing filter $W_{ij}(k)$ $(i, j = 1, 2)$ in the frequency domain, and $Y_i(k, m)$ is the T-F representation of the separated signals. Applying an inverse T-F transform to $Y_i(k, m)$, we can obtain the signals in the time domain $y_i(n)$ in this stage.

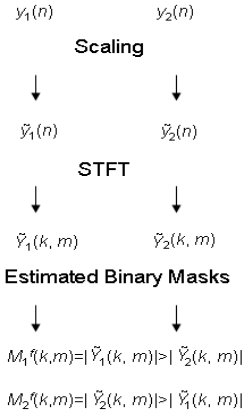


Fig. 3. Flow chart showing the second stage of the proposed method. The separated signals from the first stage i.e., $y_i(n)$ $(i = 1, 2)$ are scaled to $\tilde{y}_i(n)$, which are transformed to the T-F domain $\tilde{Y}_i(k, m)$ using the STFT. The final step is to estimate the binary masks $M_i^f(k, m)$ from $\tilde{Y}_i(k, m)$.

Sawada et al., 2004; Soon et al., 1993), are used for solving the permutation problem; however the performance of such techniques deteriorates considerably for highly reverberant acoustic conditions. To improve the quality of the separated speech signals, we consider further applying the IBM technique, as detailed in the next section.

3 Combining Convolutive ICA and Binary Masking

In order to explain the connection of this stage with the previous stage, a flow chart is shown in Figure 3. The two outputs $y_1(n)$ and $y_2(n)$ obtained from the first stage are used here to estimate the binary masks. Since these outputs are arbitrarily scaled, it is necessary to reduce the scaling ambiguity using normalisation, given as follows

$$\tilde{y}_i(n) = \frac{y_i(n)}{max(\mathbf{y}_i)} \qquad i = 1, 2 \tag{5}$$

where max denotes the maximum element of its vector argument $\mathbf{y}_i = [y_i(1), ..., y_i(L)]^T$, and L is the length of the signal. After this, we transform

the two normalized outputs into the T-F domain using the STFT

$$\tilde{Y}_i(k,m) = \mathrm{STFT}(\tilde{y}_i(n)) \qquad i = 1,2 \tag{6}$$

Without the scaling operation, the processing by (4), (5) and (6) can be omited within the algorithm. By comparing the energy of each T-F unit of the above two spectrograms, the two binary masks are estimated as (Wang, 2008)

$$M_1^f(k,m) = \begin{cases} 1 & \text{if } \mid \tilde{Y}_1(k,m) \mid > \tau \mid \tilde{Y}_2(k,m) \mid, \\ 0 & \text{otherwise} \qquad \forall k,m. \end{cases} \tag{7}$$

$$M_2^f(k,m) = \begin{cases} 1 & \text{if } \mid \tilde{Y}_2(k,m) \mid > \tau \mid \tilde{Y}_1(k,m) \mid, \\ 0 & \text{otherwise} \qquad \forall k,m. \end{cases} \tag{8}$$

where τ is a threshhold for controlling the sparseness of the mask, and $\tau = 1$ has been used in our experiment. The masks are then applied to the T-F representation of the original two-microphone recordings in order to recover the source signals, as follows

$$Y_i^f(k,m) = M_i^f(k,m) X_i(k,m) \qquad i = 1,2 \tag{9}$$

The source signals in the time domain are recovered for the purpose of pitch estimation in the next section, using the inverse STFT (ISTFT).

$$y_i^t(n) = \mathrm{ISTFT}(Y_i^f(k,m)) \qquad i = 1,2 \tag{10}$$

As observed in our experiments, the estimated IBM considerably improves the separation performance by reducing the interference to a much lower level, leading to the separated speech signals with improved quality over the outputs obtained in Section 2. However, a typical problem with the binary T-F masking is the introduction of errors in the estimation of the masks causing fluctuating musical noise (Madhu et al., 2008; Araki et al., 2005). To mitigate this problem, we employ a cepstral smoothing technique (Madhu et al., 2008) as detailed in the next section.

4 Cepstral Smoothing of the Binary Mask

The basic idea is to apply different levels of smoothing to the estimated binary mask across different frequency bands. Essentially, the levels of smoothing are determined based on the speech production mechanism. To this end, the estimated IBM is first transformed into the cepstral domain, and the different smoothing levels are then applied to the transformed mask. The smoothed mask is further converted back to the spectral domain. Through this method,

the musical artifacts within the signals can be reduced, and at the same time, the broadband structure and pitch information of the speech signal are well preserved (Madhu et al., 2008; Oppenheim and Schafer, 1975), without being noticeably affected by the smoothing operation. Representing the binary masks of (7) and (8) in the cepstrum domain we have

$$M_i^c(l,m) = DFT^{-1}\{\ln(M_i^f(k,m))\mid_{k=0,..,K-1}\} \tag{11}$$

where l and k are the quefrency bin index and the frequency bin index respectively (Madhu et al., 2008). DFT represents the discrete Fourier transform, ln denotes the natural logarithm operator and K is the length of the DFT. After applying smoothing, the resultant smoothed mask is given as

$$\overline{M}_i^s(l,m) = \gamma_l \overline{M}_i^s(l,m-1) + (1-\gamma_l)M_i^c(l,m) \quad i=1,2 \tag{12}$$

where γ_l is a parameter for controlling the smoothing level, and is selected according to the different values of l

$$\gamma_l = \begin{cases} \gamma_{env} & \text{if } l \in \{0,...,l_{env}\}, \\ \gamma_{pitch} & \text{if } l = l_{pitch}, \\ \gamma_{peak} & \text{if } l \in \{(l_{env}+1),...,K\} \setminus l_{pitch} \end{cases} \tag{13}$$

where $0 \leq \gamma_{env} < \gamma_{pitch} < \gamma_{peak} \leq 1$, l_{env} is the quefrency bin index that represents the spectral envelope of the mask $\mathbf{M}^f(k,m)$ defined as $[M_1^f(k,m), M_2^f(k,m)]^T$, and l_{pitch} is the quefrency bin index showing the structure of the pitch harmonics in $\mathbf{M}^f(k,m)$. The principle employed for this range of γ_l is illustrated as follows. $\mathbf{M}^c(l,m)$=$[M_1^c(l,m), M_2^c(l,m)]^T$, $l \in \{0,..,l_{env}\}$, basically represents the spectral envelope of the mask $\mathbf{M}^f(k,m)$. In this region the value selected for γ_l is relatively low to avoid distortion in the envelope. Similarly low smoothing is applied if l is equal to l_{pitch}, so that the harmonic structure of the signal is maintained. The symbol "\" is used to exclude l_{pitch} from the quefrency range $(l_{env}+1),...,K$. High smoothing is applied in this last range in order to reduce the artifacts without harming the pitch information and structure of the spectral envelope. Different from (Madhu et al., 2008), we calculate pitch frequency by using the segregated speech signal obtained in Section 3. Specifically pitch frequency can be computed as

$$l_{pitch} = \text{argmax}_l\{Y^c(l,m) \mid l_{low} \leq l \leq l_{high}\}, \tag{14}$$

where $Y^c(l,m)$ is the cepstrum domain representation of the segregated speech signal $y^t(n)$ obtained in (10). Note that we have omitted the subscript i in symbols γ_l, l and $Y^c(l,m)$ within (13) and (14) for notational convenience. The range l_{low}, l_{high} is chosen so that it can accommodate pitch frequencies of human speech in the range of 50 to 500 Hz. The final smoothed version of the

Table 1
The proposed multistage algorithm

1) Initialize the parameters, such as M, N, overlapfactor, and read the speech mixtures into $\boldsymbol{x}(n)$.
2) Convert $\boldsymbol{x}(n)$ to the T-F representation $\mathbf{X}(k,m)$ using STFT, and apply the constrained convolutive ICA algorithm in (Wang et al., 2005) to the mixture $\mathbf{X}(k,m)$ for estimating $\mathbf{W}(k)$. Obtain $\mathbf{Y}(k,m)$ according to (3).
3) Use (4), (5) and (6) to calculate $\tilde{Y}_i(k,m)$.
4) Estimate $M_i^f(k,m)$ according to (7) and (8), where $i=1,2$.
5) Compute $Y_i^f(k,m)$ based on (9) and $y_i^t(n)$ using (10). Compute the cepstrum domain representation of $y_i^t(n)$, i.e., $Y^c(l,m)$.
6) Calculate $M_i^c(l,m)$ in terms of (11).
7) Use (12) to calculate $\overline{M}_i^s(l,m)$, where γ_l is chosen according to (13), and $l=l_{pitch}$ is determined by (14).
8) Compute $\overline{M}_i^f(k,m)$ based on (15), and $\overline{Y}_i^f(k,m)$ according to (16).
9) Apply the ISTFT to $\overline{Y}_i^f(k,m)$ to obtain the separated signals in the time domain.

spectral mask is given as

$$\overline{M}_i^f(k,m) = exp(DFT\{\overline{M}_i^s(l,m)\mid_{l=0,\ldots,K-1}\}), \tag{15}$$

This smoothed mask is then applied to the segregated speech signals of Section 3, as follows

$$\overline{Y}_i^f(k,m) = \overline{M}_i^f(k,m)Y_i^f(k,m) \qquad i=1,2 \tag{16}$$

By further applying the ISTFT to $\overline{Y}_i^f(k,m)$, we can then obtain the separated source signals in the time domain. According to the explanation in the above sections, we summarize our algorithm in Table I.

5 Results and Comparisons

In this section, we evaluate the performance of the proposed method using simulations. The algorithm is applied to both artificially mixed signals and real room recordings.

5.1 Experimental setup and evaluation metrics

A pool of 12 different speech signals has been used in the experiments. These speech signals were uttered by six male and six female speakers with 11 different languages (Pedersen et al., 2008). All the signals have the same loudness level. The Hamming window is used with an overlap factor set to 0.75. The duration of the speech signal is 5 seconds with a sampling rate of 10 KHz. The rest of the parameters are set as: l_{env}=8, l_{low}=16, l_{high}=120, γ_{env}=0, γ_{pitch}=0.4, and γ_{peak}=0.8. Performance indices used in evaluation include signal to noise ratio (SNR), the percentage of energy loss (PEL) and the percentage of noise residue (PNR) (Hu and Wang, 2004; Pedersen et al., 2008). The expressions

of PEL and PNR are given below

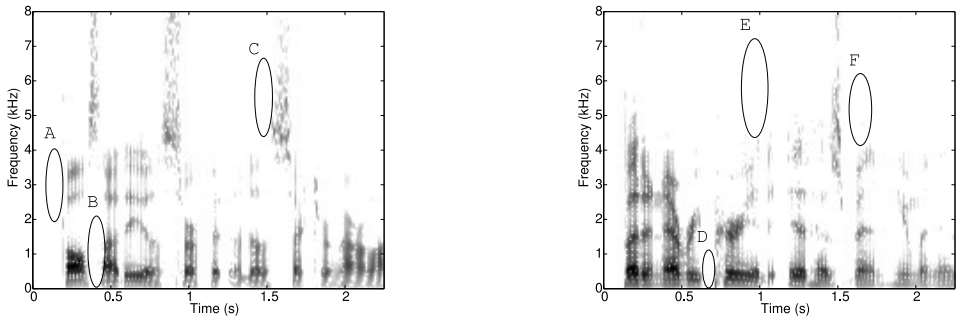
$$PEL = \frac{\sum_n (e_1^t(n))^2}{\sum_n (I^t(n))^2} \tag{17}$$

$$PNR = \frac{\sum_n (e_2^t(n))^2}{\sum_n (y^t(n))^2} \tag{18}$$

where $y^t(n)$ and $I^t(n)$ represent the estimated signal and the signal resynthesized after applying the ideal binary mask (Pedersen et al., 2008). $e_1^t(n)$ stands for the signal present in $I^t(n)$ but absent in $y^t(n)$ while $e_2^t(n)$ shows the signal present in $y^t(n)$ but absent in $I^t(n)$. SNR_i is the ratio of the desired signal to the interfering signal taken from the mixture. SNR_o is the ratio of the desired signal resynthesized from the ideal binary mask to the difference of the desired resynthesized signal and the estimated signal (Pedersen et al., 2008). Notations mSNR_i, mSNR_o and ΔSNR are also used in the evaluation where mSNR_i and mSNR_o are the average results for fifty random tests and $\Delta\text{SNR}=\text{mSNR}_o-\text{mSNR}_i$. All the SNR measurements are given in decibels (dB) in the subsequent experiments.

5.2 A separation example

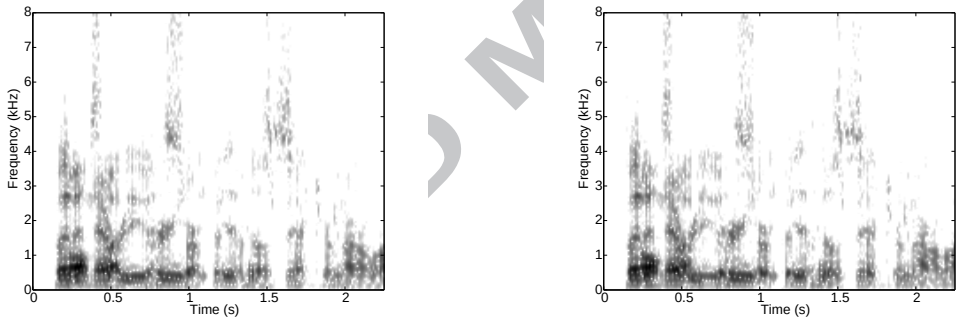
To show the performance of the proposed method for interference suppression, we present an example of applying the algorithm to the separation of two speech mixtures obtained by mixing two sources from the pool described in the above section using the simulated room model (Gaubitch, 1979), with RT set to 100 msec. The spectrograms of the two source signals are shown in Figure 4(a) and (b), and the two mixture signals in Figure 5(a) and (b). For the computation of the spectrograms, the FFT frame length was set to 2048 (i.e., 204.8 msec), and the window length (or frame shift) was fixed to 512 giving, 75% overlap between neighboring windows. Other parameters were the same as those specified in the above section. Figure 6(a) and (b) show the spectrograms of the output signals obtained from the first stage of the proposed algorithm. The results obtained from the second stage of the proposed algorithm are shown in Figure 7(a) and (b), and from the third stage in Figure 8(a) and (b). For the convenience of comparison, some T-F regions within the spectrograms are highlighted to show the performance improvement for interference suppression at each stage. In particular, we show three regions in one of the two source signals, which are marked as A, B and C for the original one (i.e. the source signal before the mixing operation) and as A_i, B_i and C_i for the separated one (i.e. the source signals estimated from the mixtures), where $i = 1, 2, 3$ is the stage index. Similarly three regions in the other source which are marked as D, E and F for the original one and as D_i, E_i and F_i for the separated one after each stage of the algorithm. From



(a) (b)

Fig. 4. Spectrograms of the two original speech signals used in the separation example. Three areas in each are highlighted for purposes of comparison with Figures 6-8.

the highlighted regions, we can observe that the interference within one source that comes from the other is reduced gradually after the processing of each stage. Compared with the output of the first stage, the interference within the estimated sources from the output of the third stage has been reduced significantly.

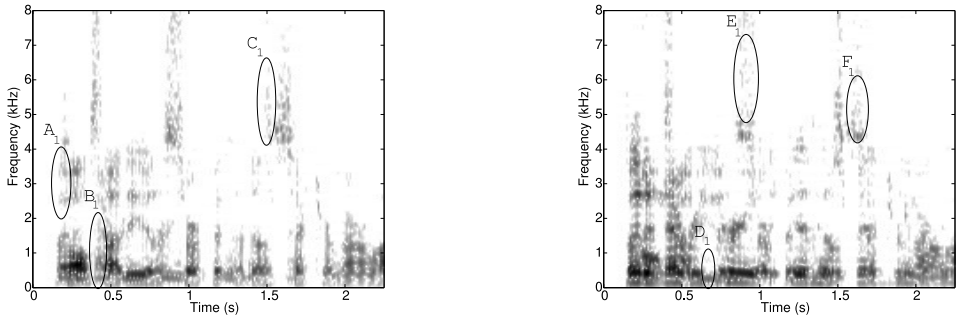


(a) (b)

Fig. 5. Spectrograms of the mixture signals that were generated by using the simulated room model with RT set to 100 msec. Both signals in (a) and (b) are the mixtures of two speech sources but with different attenuation and time delays.

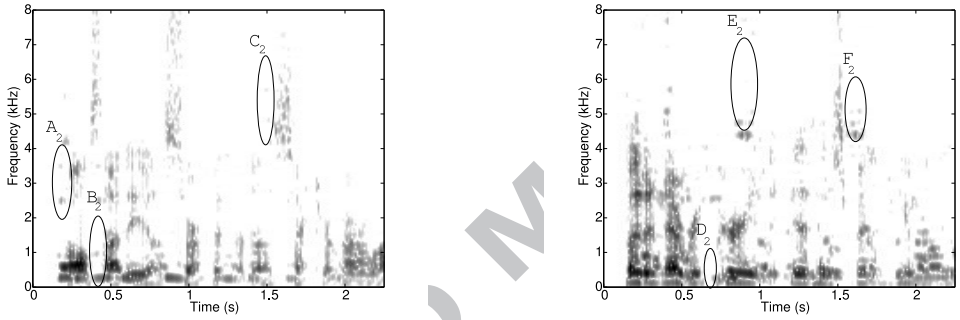
5.3 Objective evaluation

First, we evaluate the performance of the proposed algorithm for the separation of convolutive mixtures that were generated artificially by using the simulated room model (Gaubitch, 1979), for which the RT can be specified explicitly and flexibly. We wish to assess the robustness of the proposed algorithm to the changes of the key parameters used in the algorithm, such as the window length and the FFT frame length, as well as to evaluate the performance variations against different conditions for generating the mixtures,



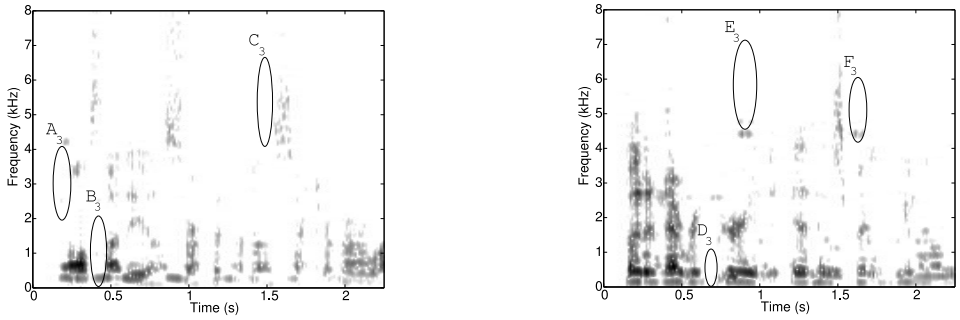
(a) (b)

Fig. 6. Spectrograms of the separated speech sources obtained from the output of the first stage of the proposed algorithm, i.e., by applying the constrained convolutive ICA algorithm. It can be observed that a considerable amount of interference from the other source still exists in the highlighted regions.



(a) (b)

Fig. 7. Spectrograms of the separated speech sources obtained from the output of the second stage of the proposed algorithm, i.e., by applying the estimated IBM. The interferences in the highlighted regions have been considerably reduced as compared with those in Figure 6.



(a) (b)

Fig. 8. Spectrograms of the separated speech sources obtained from the output of the third stage of the proposed algorithm, i.e., by applying cepstral smoothing to the estimated IBM. The interferences in the highlighted regions have been further reduced as compared with those in Figures 6 and 7.

such as the reverberation time and the noise level. In each of the subsequent experiments, we change only one parameter, i.e., the one that we intend to test, but keep all the other parameters fixed (as those already specified in Section 5.1). For each of these evaluations, the results obtained were the averaged performance of the results for 50 different convolutive mixtures, with each consisting of two speech sources randomly picked up from a pool of 12 speech signals (Pedersen et al., 2008). In the experiments, we observed that ΔSNR measured from the output of the third stage is slightly lower (hence negligible) than that measured from the output of the second stage of the proposed algorithm, although subjective listening tests suggest that the quality of the separated speech has been improved (as shown in Section 5.4). For this reason, the results of mSNR_o shown in this section are measured from the output of the second stage (as shown in our preliminary work (Jan et al., 2009)). However, more comprehensive results for mSNR_o measured at each stage of the proposed algorithm are given in Section 5.5. Analysis of variance (ANOVA) based statistical significance evaluation ((Hoel, 1976), chapter 11) of the performance difference between the second and third stage of the algorithm is also given in Section 5.5.

In the first experiment, the window length was varied from 256 to 2048 samples, while the other parameters were set identical to those in Section 5.1 and 5.2. The results are given in Table 2. It can be seen that the highest ΔSNR is obtained for the window length of 512. Therefore, the window length equal to 512 samples was used in the following experiments.

In the second experiment, the FFT frame length was changed from 512 to 2048. The average results for different FFT frame lengths are given in Table 3. It can be seen that by increasing the FFT frame length from 512 to 2048 samples, the performance of the proposed algorithm in terms of SNR, PEL and PNR is all improved. The best performance is obtained at 2048. Hence, the FFT frame length used for the subsequent experiments was fixed to 2048 samples.

In the third experiment, we change the reverberation time of the simulated room when generating the mixtures. The average results in terms of PEL, PNR and ΔSNR for the various RTs are summarized in Table 4, where the unit for RT is msec. A noticeable trend in this table is that the performance degrades gradually with an increase of RT, which is not unexpected due to the increasing sound reflections for higher room reverberations.

In the fourth experiment, we consider different levels of microphone noise by adding white noise to the mixtures, where the noise level was calculated with respect to the level of the mixtures, with a weaker noise corresponding to a smaller number (Pedersen et al., 2008). The average ΔSNR values for different noise levels are given in Table 5. It can be observed that the performance of

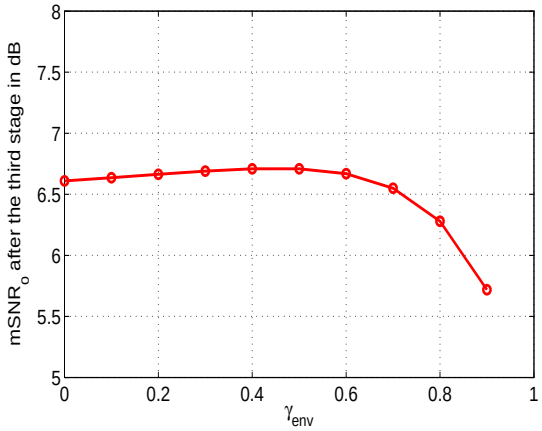


Fig. 9. Separation performance measured by mSNR$_o$ with different values of γ_{env}.

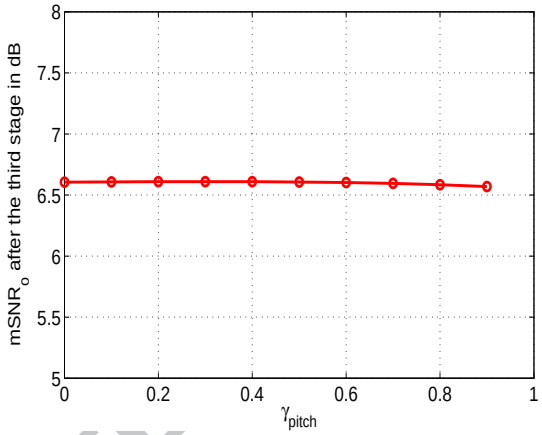


Fig. 10. Separation performance measured by mSNR$_o$ with different values of γ_{pitch}.

the algorithm decreases as the noise level is increased, and similar to (Pedersen et al., 2008), the algorithm can tolerate the noise levels up to -20 dB.

Lastly, we evaluate the performance of the proposed algorithm (without considering noise) by varying the values of γ_{env}, γ_{pitch} and γ_{peak} with the other parameters fixed as: RT =100 msec, window length=512, and NFFT=2048. The values of γ_{env} , γ_{pitch} and γ_{peak} as discussed in section 4, were chosen in the range $[0, 0.9]$. The results measured by mSNR$_o$ are given in Figures 9, 10 and 11 respectively. From Figure 9, it is observed that mSNR$_o$ after the third stage increases slowly for γ_{env} ranging from 0 to 0.4 and then starts decreasing. Figure 10 shows a very slight increase in mSNR$_o$ when γ_{pitch} is between 0 and 0.5 followed by a very slight decrease. In Figure 11, mSNR$_o$ first increases slowly when γ_{peak} varies from 0 to 0.4 and then a sharp decrease is observed when γ_{peak} is between 0.5 and 0.9. These experiments show that the separation performance varies to some extent when different values for γ_{env}, γ_{pitch} and γ_{peak} are used.

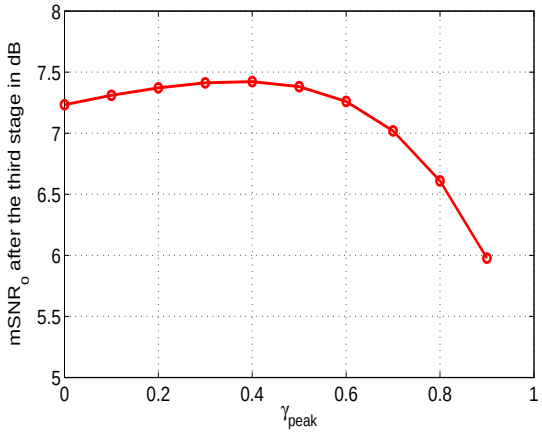


Fig. 11. Separation performance measured by mSNR_o with different values of γ_{peak}.

Table 2
Separation Results for Different Window Lengths

Window Length	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
256	9.10	15.30	1.10	7.11	6.01
512	8.60	14.48	1.10	7.44	6.34
1024	9.30	14.70	1.10	7.11	6.01
2048	10.92	15.92	1.12	6.32	5.20

Table 3
Separation Results for Different FFT Frame lengths

NFFT	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
512	9.06	14.96	1.10	7.17	6.06
1024	8.65	14.53	1.10	7.40	6.30
2048	8.60	14.48	1.10	7.44	6.34

Table 4
Separation Results for Different *RT*

RT	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
40	2.16	2.24	1.13	13.22	12.08
60	3.79	4.12	1.15	10.94	9.79
80	5.50	8.30	1.14	9.42	8.27
100	8.60	14.48	1.10	7.44	6.34
120	10.99	19.53	1.03	6.30	5.26
140	13.36	24.14	0.94	5.48	4.53
150	13.86	25.38	0.90	5.29	4.39

5.4 Listening tests

As mentioned in the above section that ΔSNR measured from the output of the third stage of the proposed algorithm appears to be slightly lower than that measured from the output of the second stage of the proposed algorithm (see more results and detailed analysis in the next section). This suggests that

Table 5
Separation Results for Different Noise Levels

Noise	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
-40 dB	8.60	14.48	1.10	7.45	6.34
-30 dB	8.60	14.48	1.10	7.44	6.34
-20 dB	8.62	14.52	1.10	7.43	6.33
-10 dB	9.46	16.49	1.09	6.91	5.81

Table 6
MOS Obtained From Subjective Listening Tests

RT	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
50	3.26	3.90	3.01	5.0948	4.1960	0.0320
100	2.12	2.62	2.29	4.7094	4.1960	0.0386
150	1.87	2.39	2.02	5.0995	4.1960	0.0319
200	1.09	2.07	1.82	50.2059	4.1960	0.0000

cepstral smoothing actually does not improve the objective performance in terms of SNR measurement (see also (Wang, 2008)). Nevertheless, our informal listening tests seem to contradict the SNR measurements and confirm that the cepstral smoothing does improve the quality of the separated speech, especially for the musical noise removal. To show this, we conducted subjective listening tests by recruiting 15 participants with normal hearing. Each of these listeners was asked to give an integer score ranging from 1 (musical noise clearly audible) to 5 (noise not audible) for the final segregated speech signals, as suggested in (Araki et al., 2005). During these tests, each participant was asked to listen to 2 groups of separated speech signals obtained in the experiments where RT was set to 50, 100, 150 and 200 msec respectively, with one group containing y_1 and the other group containing y_2. A total of 8 groups of speech signals were evaluated subjectively by these participants. Each group was composed of 3 speech signals, i.e. the estimated source obtained from the output of the second stage, the one from the third stage, and the source signal estimated by Pedersen *et al.*'s method. Note that the listeners had no prior knowledge on which signal was obtained from which algorithm. This ensures a fair comparison between the algorithms. The mixtures used in these tests were generated by the simulated room model with RT equal to 50, 100, 150 and 200 msec, respectively. The scores given by the listener are provided on the basis of how clean the separated signals from the two stages are in comparison to each other, or how much musical noise is present in the separated signals. A signal with less musical noise is cleaner, and hence is given a higher mean opinion score (MOS) (Araki et al., 2005). The average results of MOS for the 15 listeners are given in Table 6. It indicates that using cepstral smoothing gives higher MOS, suggesting the improved quality of the separated speech. To examine whether the improvement in MOS after smoothing is statistically significant, we perform one-way ANOVA based F-test (Hoel, 1976) for the MOS obtained before and after smoothing. The results are given in Table 6.

Table 7
MOS Obtained From Subjective Listening Tests For Different Window Lengths

For RT=100 msec						
Window Length	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
256	2.35	3.70	2.57	64.4233	4.0980	0.00000
512	2.70	3.65	2.90	16.5277	4.0980	0.00023
1024	2.60	3.65	2.81	24.1470	4.0980	0.00001
2048	2.40	3.10	2.64	7.0000	4.0980	0.0118
For RT=200 msec						
Window Length	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
256	1.70	2.80	1.94	16.7810	4.0980	0.00021
512	1.75	2.70	2.04	21.5016	4.0980	0.00004
1024	1.75	2.65	2.01	15.1626	4.0980	0.00038
2048	1.55	2.35	1.78	15.6903	4.0980	0.00031

Table 8
MOS Obtained From Subjective Listening Tests For Different FFT Frame Lengths

For RT=100 msec						
NFFT	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
512	3.30	4.10	2.88	17.3714	4.0980	0.00017
1024	3.20	4.15	2.87	17.3646	4.0980	0.00017
2048	2.70	3.65	2.90	16.5277	4.0980	0.00023
For RT=200 msec						
NFFT	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
512	2.05	2.80	1.89	8.8509	4.0980	0.00510
1024	1.75	2.50	1.96	10.3012	4.0980	0.00270
2048	1.75	2.70	2.04	21.5016	4.0980	0.00004

The critical value (F_{crit}) is the number that the test statistic must overcome to reject the test. The p-value stands for the probability of a more extreme (positive or negative) result than what we actually achieved, given that the null hypothesis is true. F-value can be defined as the ratio of the variance of the group means to the mean of the within group variances. All the F-tests in this work have been carried out at 5% significance level. If $F < F_{crit}$ and p-value is greater than 0.05 (5% significance level), then the given results are statistically insignificant. It can be observed that the p-values obtained for all the cases of RT in Table 6 are smaller than 0.05, suggesting that the improvement in all the four cases is statistically significant.

Table 9
MOS Obtained From Subjective Listening Tests For Different Noise Levels

For RT=100 msec						
Noise	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
-40 dB	3.30	4.20	2.84	15.8660	4.0980	0.00029
-30 dB	3.20	4.15	2.70	19.3211	4.0980	0.00008
-20 dB	2.70	3.70	2.09	14.3939	4.0980	0.00051
-10 dB	1.80	2.55	1.84	10.6079	4.0980	0.00240
For RT=200 msec						
Noise	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
-40 dB	2.00	2.80	2.01	16.0000	4.0980	0.00028
-30 dB	2.15	2.85	1.93	12.3311	4.0980	0.00120
-20 dB	1.70	2.50	1.76	18.4242	4.0980	0.00011
-10 dB	1.30	1.90	1.49	9.7714	4.0980	0.0034

Additional listening tests have been carried out using the speech signals randomly selected from the experimental results employed for the objective evaluation of the proposed method. We have recruited 20 volunteers to participate the subjective listening tests, including the 15 listeners mentioned earlier. The results have been evaluated for different window lengths in Table 7, for different FFT frame lengths in Table 8 and for different noise levels in Table 9. The RT has been set to 100 and 200 msec, respectively. The criteria used in Table 6 for the MOS have also been employed here. The results given in Table 7 show that for different window lengths at RT = 100 and 200 msec, cepstral smoothing offers higher MOS scores, indicating that the quality of the segregated speech signal has been improved. A similar trend can be observed in Table 8 and 9 where using cepstral smoothing achieves a higher MOS. In all cases the differences of MOS before and after smoothing are statistically significant.

5.5 Comparison to other methods

In this section, we compare the proposed multistage method with two related approaches in (Pedersen et al., 2008) and (Wang et al., 2005). In (Wang et al., 2005) speech signals were separated from convolutive mixtures by exploiting the second order non-stationarity of the sources in the frequency domain, where the cross-power spectrum based cost function and a penalty function have been employed to convert the separation problem into a joint diagonalization problem with unconstrained optimization. Pedersen *et al.*'s method combines an instantaneous ICA algorithm with the binary T-F masking for

Table 10
Comparison results for Different Window Lengths

Window Length	$mSNR_i$	$mSNR_o$ after the 1st stage	$mSNR_o$ after the 2nd stage	$mSNR_o$ after the 3rd stage	ANOVA test for the difference between the SNR_os from the 2nd and 3rd stage		
					F-value	F_{crit}	p-value
256	1.10	2.98	7.11	6.81	0.9085	3.9380	0.3429
512	1.10	3.02	7.44	6.59	7.6412	3.9380	0.0068
1024	1.10	3.01	7.11	6.09	11.4642	3.9380	0.0010
2048	1.12	2.95	6.32	5.32	12.8289	3.9380	0.0005

Table 11
Comparison results for Different FFT Frame Lengths

NFFT	$mSNR_i$	$mSNR_o$ after the 1st stage	$mSNR_o$ after the 2nd stage	$mSNR_o$ after the 3rd stage	ANOVA test for the difference between the SNR_os from the 2nd and 3rd stage		
					F-value	F_{crit}	p-value
512	1.10	3.01	7.17	6.46	5.8298	3.9380	0.0176
1024	1.10	3.02	7.40	6.57	7.4946	3.9380	0.0074
2048	1.10	3.02	7.44	6.59	7.6412	3.9380	0.0068

underdetermined blind source separation, where the outputs of the ICA algorithm were used to estimate the binary mask in an iterative way to extract multiple speech sources from two mixtures.

Comparison between the proposed method and the method in (Wang et al., 2005) is essentially equivalent to the comparison between the outputs from the third (and/or second stage) and those from the first stage, as the method in (Wang et al., 2005) is employed in the first stage of the proposed approach. Therefore, without performing additional experiments, we show more results that were obtained from the experiments already conducted in Section 5.3. In parallel with the results shown in Tables 2, 3, 4, and 5, we show the comparison results in terms of $mSNR_o$ in Tables 10 for different window lengths, 11 for different FFT frame lengths, 12 for different RT values and 13 for different noise levels. All the results were measured based on 50 random tests. Note that $mSNR_o$ obtained after the first stage of the proposed method is approximately calculated. This is because, according to the definition of SNR_o in Section 5.1, the masked output signals should be used for the calculation of output SNR, while the obtained signal from the output of the first stage (Wang et al., 2005) is not a masked signal. The results in Table 10 clearly indicate that the output SNR has been improved at the second and third stage in comparison to the first stage for different window lengths. The objective results from the third stage in terms of $mSNR_o$ measurement are slightly worse than those of the second stage, due to the smoothing operation. According to our subjective listening tests in the previous section, the quality of the speech source from the third stage is actually improved, due to the reduced level of audible musical noise.

Table 12
Comparison results for Different RT

RT	mSNR$_i$	mSNR$_o$ after the 1st stage	mSNR$_o$ after the 2nd stage	mSNR$_o$ after the 3rd stage	ANOVA test for the difference between the SNR$_o$s from the 2nd and 3rd stage		
					F-value	F$_{\text{crit}}$	p-value
40	1.13	3.70	13.22	9.44	100.2190	3.9380	0.0000
60	1.15	3.47	10.94	8.48	40.4630	3.9380	0.0000
80	1.14	3.36	9.42	7.75	23.1972	3.9380	0.0000
100	1.10	3.02	7.44	6.59	7.6412	3.9380	0.0068
120	1.03	2.70	6.30	5.82	3.7015	3.9380	0.0573
140	0.94	2.47	5.48	5.23	0.9266	3.9380	0.3381
150	0.90	2.42	5.29	5.11	0.5210	3.9380	0.4721

Table 13
Comparison results for Different Noise Levels

Noise	mSNR$_i$	mSNR$_o$ after the 1st stage	mSNR$_o$ after the 2nd stage	mSNR$_o$ after the 3rd stage	ANOVA test for the difference between the SNR$_o$s from the 2nd and 3rd stage		
					F-value	F$_{\text{crit}}$	p-value
-40 dB	1.10	3.02	7.45	6.60	7.6297	3.9380	0.0069
-30 dB	1.10	3.02	7.44	6.60	7.6186	3.9380	0.0069
-20 dB	1.10	3.02	7.43	6.59	7.5950	3.9380	0.0070
-10 dB	1.09	3.06	6.91	6.09	8.2232	3.9380	0.0051

Table 11 compares the results of the proposed method and the method in (Wang et al., 2005) for different FFT frame lengths, where the window length was fixed to 512, the overlap factor and RT remained the same as those used for Table 10. From this table, we can also observe the improved performance of the proposed method in terms of SNR measurements, as compared with the method in (Wang et al., 2005). Subjective listening tests also show that our results have considerably improved quality over those in (Wang et al., 2005) for different FFT frame lengths, which are consistent with the SNR measurements. In Table 12, comparison has been made for different values of RT, where the window length and the overlap factor were identical to those used in Table 11, and the FFT frame length was the same as that in 10. The results show that the output SNR decreases with an increase in RT, and the proposed method has better performance in terms of the averaged output SNR. Specifically, when RT equals to 100 msec, mSNR$_o$ of the third stage is approximately 4 dB higher than that of the first stage. The improvement is more prominant when RT is relatively low. In Table 13 we performed experiments by considering the microphone noise in the mixture, as discussed already in Table 5. In this table, RT was set to 100 msec, and other parameters were the same as those in Table 12. It can be observed that the proposed method performs better than the method in (Wang et al., 2005) for the separation of noisy mixtures. Specifically, comparing mSNR$_o$ between the first and third stages, we see that there is about 3 dB improvement for noise level at -10 dB, and 3.6 dB for noise

Table 14
Comparison of Separation Performance and Computational Cost Between the Proposed Method and Pedersen Et AL.'s method

Algorithm	PEL	PNR	ΔSNR	Total time	Time per test	Run time memory requirement [2]
Proposed	30.56	9.73	2.50	40min	0.8min	223.28 MB
Pedersen *et al.*	17.14	49.33	2.64	700min	14min	255.17 MB

[2]Note that the results also include the memory required for the matlab software

level at -30 dB. The results discussed above show that our proposed method outperforms the method in (Wang et al., 2005) in terms of SNR measurements.

To determine whether the relatively small differences of mSNR_o between the second and third stage of the proposed method are statistically significant, we perform one-way ANOVA based F-test (Hoel, 1976) as described in Section 5.4. The testing results are given in Table 10, 11, 12 and 13. To explain how the F-test was applied to the results, we take the case of NFFT equal to 512 (in Table XI) as an example, where mSNR_o after the second and third stage is 7.17 dB and 6.46 dB respectively. Both mSNR_os were calculated by averaging 50 individual SNR_os obtained from the 50 random tests. Each group of 50 SNR_os forms a vector, and hence two vectors can be formed from the second and third stage. The F-value was then computed from these two vectors, which is 5.8298. The F-values in other cases and tables were computed in the same way. From the results in these tables, we can observe that in many testing cases the differences of mSNR_o between the second and third stage of the proposed algorithm, although small, are statistically significant whereas in some cases the differences are insignificant.

The performance of the proposed method is also compared with the algorithm in (Pedersen et al., 2008) in terms of both computational complexity and separation quality. The separation quality is measured objectively using SNR measurement as in the above experiments, and subjectively by listening tests. To make this comparison, we use the real room recordings that were obtained in (Pedersen et al., 2008). The real recordings were made in a reverberant room with $RT = 400$ msec. Two omnidirectional microphones vertically placed and closely spaced are used for the recordings. Different loudspeaker positions are used to measure the room impulse responses. Details about the recordings can be found in (Pedersen et al., 2008) and are not given here. Clean speech signals from the pool of 12 speakers were convolved with the room impulses to generate the source signals (Pedersen et al., 2008). The specifications of the computing facilities that were used to perform the experiments include Intel(R) Xeon(TM) 3.00GHz CPU and 31.48 GB memory. The results are given in Table 14. The results show that our proposed algorithm is 18 times faster than the Pedersen et al. method. Their method requires 700 minutes for 50 random tests and 14 minutes per test. In contrast our proposed method is much faster and requires 40 minutes for 50 tests and 0.8 minutes per test. The

time computational complexity of both methods was also approximately calculated. The order of complexity of our proposed method is $O(I_3(MFKlogK + M)) + O(I_3KMN(2N + M)) + O(MNI_3K) + O(FKlogK) + O(NKF) + O(L)$, where F is the number of frames [3], L is the length of the signal, and I_3 denotes the required number of iterations for the constrained convolutive ICA algorithm (Wang et al., 2005) to converge. Similarly the complexity of the Pedersen et al. method is $O(FKlogKI_2) + O(NKFI_2) + O(NMI_1I_2)$, where I_1 is the iteration number for the INFORMAX algorithm (used as a first stage in their method) to converge, while I_2 denotes the total number of iterations for the Pedersen et al. method to segregate the speech mixtures. Although the results for ΔSNR are comparable, listening tests given in Table 6 suggest that our results have a better quality than those in (Pedersen et al., 2008). Some demos are available on the website (Wang, 2010) for both real and artificial recordings.

6 Conclusion

The proposed approach consists of three major steps. A convolutive ICA algorithm (Wang et al., 2005) is first applied in order to take into account the reverberant mixing environments based on a convolutive unmixing model. Binary T-F masking is used in the second step for improving the SNR of the separated speech signal, due to its effectiveness in rejecting the energy of interference by assigning zeros to the T-F units in the masking matrix in which the energy of the interference is stronger than the target speech. The artifacts (musical noise) due to the error in the estimation of the binary mask in the segregated speech signals are further reduced by applying the cepstral smoothing technique. Compared with smoothing directly in the spectral domain, cesptral smoothing has the advantage of preserving the harmonic structure of the separated speech signal while reducing the musical noise to a lower level by smoothing out the unwanted isolated random peaks.

In comparison to (Wang et al., 2005), the considerable improvement achieved by the proposed method in terms of both objective measurements using SNR and subjective listening tests is mainly due to the introduction of the binary T-F masking operation and the cepstral smoothing. The binary masking contributed mostly to the improvement of interference cancellation, and cepstral smoothing further improves the perceptual quality of the separated speech. For a reverberation time of 100 msec, the proposed algorithm achieves approximately 4 dB SNR gain over a typical convolutive ICA algorithm in (Wang et al., 2005). Compared with (Wang et al., 2005), the computational complexity of the proposed algorithm is higher due to the additional processing of

[3] If there is no overlap between adjacent frames then $F \cdot K \approx L$.

IBM and cepstral smoothing. It is however still computationally efficient as FFT and its inverse are used for the transforms in all the steps.

Note the difference between our proposed method and Pedersen *et al.*'s method (Pedersen et al., 2008) despite a similar combination of an ICA algorithm with the IBM technique. First, our proposed algorithm directly addresses the convolutive BSS model based on the frequency-domain approach, while Pedersen *et al.*'s method is based on an instantaneous model and an instantaneous ICA algorithm, even though their algorithm has also been tested for convolutive mixtures. Second, the algorithm in (Pedersen et al., 2008) is iterative, which is computationally demanding. Moreover, we have introduced cepstral smoothing, which has the advantage of reducing the musical artifacts caused by the IBM technique.

In future work we plan to extend the proposed algorithm to underdetermined cases. Another important issue is how to deal with highly reverberant speech mixtures. One could analyze reverberation effects and reduce such effects present in the microphone signal before applying the ICA and IBM approaches. This issue will be addressed in our subsequent research.

Acknowledgments

We are grateful to M. S. Pedersen for providing the matlab code of (Pedersen et al., 2008) and the assistance in the preparation of this work. Part of the work was conducted while W. Wang was visiting OSU. T. U. Jan was supported by the NWFP UET Peshawar, Pakistan. W. Wang was supported in part by a Royal Academy of Engineering travel grant (IJB/AM/08-587) and an EPSRC grant (EP/H012842/1). D. L. Wang was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707).

References

Aissa-El-Bey, A., Abed-Meraim, K., Grenier, Y., July 2007. Blind separation of underdetermined convolutive mixtures using their time-frequency representation. IEEE Trans. on Audio, Speech, and Lang. Process. 15, 1540–1550.

Amari, S., Douglas, S. C., Cichocki, A., Wang, H. H., 1997. Multichannel blind deconvolution and equalization using the natural gradient. In: IEEE Workshop on Signal Process. pp. 101–104.

Araki, S., Makino, S., Sawada, H., Mukai, R., 2004. Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica. In: Proc. 5th Int. Conf. Independent Component Anal. and Blind Signal Separation. Granada, Spain, pp. 898–905.

Araki, S., Makino, S., Sawada, H., Mukai, R., 2005. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In: Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. Vol. 3. USA, pp. 81–84.

Araki, S., Mukai, R., Makino, S., Saruwatari, H., 2003. The fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. IEEE Trans. on Speech and Audio Process. 11, 109–116.

Araki, S., Sawada, H., Mukai, R., Makino, S., 2007. Underdetermined blind sparse source separation for arbitrarily arranged multiple sources. EURASIP J. App. Signal Process. 87, 1833–1847.

Back, A. D., Tosi, A. C., 1994. Blind deconvolution of signals using a complex recurrent network. In: IEEE Workshop Neural Netw. Signal Process. pp. 565–574.

Buchner, H., Aichner, R., Kellermann, W., 2004. Audio Signal Processing for Next-Generation Multimedia Communication Systems. Kluwer Academic Publishers, Boston/Dordrecht/London, Ch. Blind source separation for convolutive mixtures: A unified treatment, In Y.Huang and J. Benesty (eds.), pp. 255–293.

Cherry, E. C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Amer. 25 (5), 975–979.

Cichocki, A., Amari, S., 2002. Adaptive Blind Signal and Image Processing. Wiley Press.

Dillon, H., 2001. Hearing aids. New York: Thieme.

Douglas, S., Sawada, H., Makino, S., January 2005. Natural gradient multichannel blind deconvolution and speech separation using causal fir filters. IEEE Trans. on Speech Audio Process. 13 (1), 92–104.

Douglas, S. C., Gupta, M., Sawada, H., Makino, S., 2007. Spatio-temporal fastica algorithms for the blind separation of convolutive mixtures. IEEE Trans. on Audio, Speech, and Language Process. 15 (5), 1511–1520.

Douglas, S. C., Sun, X., December 2002. Convolutive blind separation of speech mixtures using the natural gradient. Speech Communication 39, 65–78.

Gaubitch, N. D., 1979. Allen and berkley image model for room impulse response, imperial college london [online]. Available: http://www.commsp.ee.ic.ac.uk/%7Endg/downloadfiles/mcsroom.m.

Han, S., Cui, J., Li, P., 2009. Post-processing for frequency-domain blind source separation in hearing aids. In: Proc. 7th Int. Conf. on Information, Communications and Signal Process., ICICS. pp. 356–360.

He, Z., Xie, S., Ding, S., Cichocki, A., 2007. Convolutive blind source separation in the frequency domain based on sparse representation. IEEE Trans. on Audio, Speech, and Lang. Process. 15, 1551–1563.

Hoel, P. G., 1976. Elementary Statistics, 4th Edition. New York: Wiley.

Hu, G., Wang, D. L., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Netw. 15, 1135–1150.

Hyvarinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley and Sons.

Jan, T. U., Wang, W., Wang, D. L., April 2009. A multistage approach for blind separation of convolutive speech mixtures. In: IEEE Int. Conf. on Acoustics, Speech and Signal Process. Taiwan, pp. 1713–1716.

Lambert, R. H., Bell, A. J., April 1997. Blind separation of multiple speakers in a multipath environment. In: IEEE Int. Conf. on Acoustics, Speech and Signal Process. pp. 423–426.

Lee, T. W., September 1998. Independent Component Analysis: Theory and Applications. Kluwer Academic Publishers.

Lee, T. W., Bell, A. J., Orglmeister, R., June 1997. Blind source separation of real world signals. In: IEEE Int. Conf. on Neural Netw. pp. 2129–2135.

Madhu, N., Breithaupt, C., Martin, R., 2008. Temporal smoothing of spectral masks in the cepstral domain for speech separation. In: Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. Las Vegas, USA, pp. 45–48.

Makino, S., Sawada, H., Mukai, R., Araki, S., July 2005. Blind source separation of convolutive mixtures of speech in frequency domain. IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences E88-A (7), 1640–1655.

Matsuoka, K., Nakashima, S., December 2001. Minimal distortion principle for blind source separation. In: Int. Conf. Independent Component Anal. San Diego, CA, USA, pp. 722–727.

Mazur, R., Mertins, A., 2009. An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models. IEEE Trans. on Audio, Speech, and Lang. Process. 17, 117–126.

Mitianondis, N., Davies, M., 2002. Audio source separation: solutions and problems. Int. J. Adaptive Control and Signal Process., 1–6.

Mukai, R., Sawada, H., Araki, S., Makino, S., September 2004. Frequency domain blind source separation for many speech signals. In: Int. Conf. on Independent Component Anal. pp. 461–469.

Murata, N., Ikeda, S., Ziehe, A., October 2001. An approach to blind source separation based on temporal structure of speech signals. Neurocomputing 41 (1-4), 1–24.

Nesta, F., Omologo, M., Svaizer, P., 2008. Separating short signals in highly reverberant environment by a recursive frequency-domain bss. In: Proc. Hands-Free Speech Communication and Microphone Arrays, HSCMA. Trento, Italy, pp. 232–235.

Nesta, F., Wada, T. S., Juang, B. H., 2009. Coherent spectral estimation for a robust solution of the permutation problem. In: IEEE Workshop on Applications of Signal Process. to Audio and Acoustics. pp. 105–108.

Nickel, R. M., Iyer, A. N., 2006. A novel approach to automated source separation in multispeaker environments. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Process. pp. 629–632.

Olsson, R. K., Hansen, L. K., 2006. Blind separation of more sources than sensors in convolutive mixtures. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Process. pp. 657–660.

Oppenheim, A. V., Schafer, R. W., 1975. Digital Signal Processing. Prentice

Hall, New Jersey.

Parra, L., Spence, C., May 2000. Convolutive blind separation of non stationary sources. IEEE Trans. on Speech Audio Process. 8, 320–327.

Pedersen, M. S., Wang, D. L., Larsen, J., Kjems, U., March 2008. Two-microphone separation of speech mixtures. IEEE Trans. on Neural Netw. 19, 475–492.

Rahbar, K., Reilly, J. P., 2005. A frequency domain method for blind source separation of convolutive audio mixtures. IEEE Trans. on Speech and Audio Process. 13 (5), 832–844.

Reju, V. G., Koh, S. N., Soon, I. Y., 2010. Underdetermined convolutive blind source separation via time-frequency masking. IEEE Trans. on Audio, Speech, and Lang. Process. 18, 101–116.

Rodrigues, G. F., Yehia, H. C., 2009. Limitations of the spectrum masking technique for blind source separation. In: Proc. 8th Independent Component Anal. and Signal Separation. pp. 621–628.

Roman, N., Wang, D. L., Brown, G. J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Amer. 114, 2236–2252.

Sawada, H., Araki, S., Mukai, R., Makino, S., 2006. Blind extraction of dominant target sources using ica and time-frequency masking. IEEE Trans. on Audio, Speech, and Lang. Process. 14 (6), 2165–2173.

Sawada, H., Araki, S., Mukai, R., Makino, S., 2007. Grouping separated frequency components by estimating propagation model parameters in frequency domain blind source separation. IEEE Trans. on Audio, Speech, and Lang. Process. 15, 1592–1604.

Sawada, H., Mukai, R., Araki, S., Makino, S., March 2003. Polar coordinate based nonlinear function for frequency-domain blind source separation. IEICE Trans. Fundamentals E86-A (3), 590–596.

Sawada, H., Mukai, R., Araki, S., Makino, S., September 2004. A robust and precise method for solving the permutation problem of frequency domain blind source separation. IEEE Trans. on Speech Audio Process. 12, 530–538.

Schobben, L., Sommen, W., August 2002. A frequency domain blind signal separation method based on decorrelation. IEEE Trans. on Signal Process. 50 (8), 1855–1865.

Smaragdis, P., 1998. Blind separation of convolved mixtures in the frequency domain. Neurocomputing, 21–34.

Soon, V. C., Tong, L., Huang, Y. F., Liu, R., 1993. A robust method for wideband signal separation,. In: Proc. IEEE Int. Symp. Circuits Syst. Vol. 1. Chicago, USA, pp. 703–706.

Wang, D. L., 2005. Speech Separation by Humans and Machines. Kluwer Academic, Norwell MA, Ch. On ideal binary mask as the computational goal of auditory scene analysis, pp. 181–297.

Wang, D. L., 2008. Time-frequency masking for speech separation and its potential for hearing aid design. Trends in Amplification 12, 332–353.

Wang, D. L., Brown, G. J., 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley/IEEE Press, Hoboken NJ.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., Lunner, T., 2009. Speech intelligibility in background noise with ideal binary time-frequency masking. J. Acoust. Soc. Amer. 125, 2336–2347.

Wang, W., 2010. Available: http://personal.ee.surrey.ac.uk/Personal/W.Wang/demondata.html.

Wang, W., Chambers, J. A., Sanei, S., 2004. A novel hybrid approach to the permutation problem of frequency domain blind source separation. In: Proc. 5th Int. Conf. on Independent Component Anal. and Blind Signal Separation. Granada, Spain, pp. 530–537.

Wang, W., Sanei, S., Chambers, J. A., May 2005. Penalty function-based joint diagnolization approach for convolutive blind separation of nonstationary sources. IEEE Trans. on Signal Process. 53, 1654–1669.

Yoshioka, T., Nakatani, T., Miyoshi, M., 2009. Fast algorithm for conditional separation and dereverberation. In: Proc. of the 17th European Signal Process. Conf. pp. 1432–1436.