Two-microphone Separation of Speech Mixtures

Michael Syskind Pedersen, *Member, IEEE,* DeLiang Wang, *Fellow, IEEE,* Jan Larsen, *Senior Member, IEEE,* and Ulrik Kjems, *Member, IEEE*

***Abstract*—Separation of speech mixtures, often referred to as the cocktail party problem, has been studied for decades. In many source separation tasks, the separation method is limited by the assumption of at least as many sensors as sources. Further, many methods require that the number of signals within the recorded mixtures be known in advance. In many real-world applications these limitations are too restrictive. We propose a novel method for underdetermined blind source separation using an instantaneous mixing model which assumes closely spaced microphones. Two source separation techniques have been combined, *independent component analysis* (ICA) and *binary time-frequency masking*. By estimating binary masks from the outputs of an ICA algorithm, it is possible in an *iterative* way to extract basis speech signals from a convolutive mixture. The basis signals are afterwards improved by grouping similar signals. Using two microphones we can separate in principle an arbitrary number of mixed speech signals. We show separation results for mixtures with as many as seven speech signals under instantaneous conditions. We also show that the proposed method is applicable to segregate speech signals under reverberant conditions, and we compare our proposed method to another state-of-the-art algorithm. The number of source signals is not assumed to be known in advance and it is possible to maintain the extracted signals as stereo signals.**

***Index Terms*—Underdetermined speech separation, ICA, time-frequency masking, ideal binary mask.**

I. Introduction

THE problem of extracting a single speaker from a mixture of many speakers is often referred to as the cocktail party problem [1], [2]. Human listeners cope remarkably well in adverse environments, but when the noise level is exceedingly high, human speech intelligibility also suffers. By extracting speech sources from a mixture of many speakers, we can potentially increase the intelligibility of each source by listening to the separated sources.

Blind source separation addresses the problem of recovering N unknown source signals $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ from M recorded mixtures $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ of the source signals. n denotes the discrete time index. Each of the recorded mixtures $x_i = x_i(n)$ consists of $N_s = f_s T$ samples, where f_s is the sampling frequency and T denotes the duration in seconds. The term 'blind' refers to that only the recorded mixtures are known. The mixture is assumed to be a linear superposition of the source signals, sometimes with additional noise, i.e.,

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

Michael Syskind Pedersen and Ulrik Kjems are with Oticon A/S, Kongebakken 9, DK-2765, Denmark. Email: {msp, uk}@oticon.dk, DeLiang Wang is with the Department of Computer Science & Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277, U.S.A. Email: dwang@cse.ohio-state.edu. Jan Larsen is with the Intelligent Signal Processing Group at the Department of Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark. Email: jl@imm.dtu.dk.

where $\mathbf{A}$ is an $M \times N$ mixing matrix. $\boldsymbol{\nu}(n)$ is additional noise. Also, $\mathbf{A}$ is assumed not to vary as function of time. Often, the objective is to estimate one or all of the source signals. An estimate $\mathbf{y}(n)$ of the original source signals can be found by applying an (pseudo) inverse linear operation, i.e.,

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n), \tag{2}$$

where $\mathbf{W}$ is an $N \times M$ separation matrix. Notice that this inversion is not exact when noise is included in the mixing model. When noise is included as in (1), $\mathbf{y}(n)$ is a nonlinear function of $\mathbf{x}(n)$ [3]. In this paper, the inverse is approximated by a linear system.

In real environments, a speech signal does not only arrive from a single direction. Rather, multiple reflections from the surroundings occur as delayed and filtered versions of the source signal. In this situation, the mixing model is better approximated by a *convolutive* mixing model. The convolutive FIR mixture is given as

$$\mathbf{x}(n) \quad = \quad \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{s}(n-k) + \boldsymbol{\nu}(n) \tag{3}$$

Here, the source signals are mixtures of filtered versions of the anechoic source signals. The filters are assumed to be causal and of finite length K. Numerous algorithms have been proposed to solve the convolutive problem [4], but few are able to cope with underdetermined as well as reverberant conditions [5]–[9].

Independent Component Analysis (ICA) describes a class of methods that retrieve the original signals up to an arbitrary permutation and scaling [10]. Successful separation relies on assumptions on the statistical properties of the source signals. To obtain separation, many ICA methods require that at most one source be Gaussian. Many algorithms assume that the source signals are independent or the source signals are non-Gaussian [11]–[14]. Other methods are able to separate the source signals using only second order statistics. Here, it is typically assumed that the sources have different correlation [15]–[17] or the source signals are non-stationary [18], [19]. Blind source separation algorithms have been applied in many areas such as feature extraction, brain imaging, telecommunications, and audio separation [10].

ICA methods have several drawbacks. Often, it is required that the number of source signals is known in advance and only few have addressed the problem of determining the number of sources in a mixture [20], [21]. Further, standard formulation requires that the number of source signals does not exceed the number of microphones. If the number of sources is greater than the number of mixtures, the mixture is called *underdetermined* (or *overcomplete*). In this case, the independent

components cannot be recovered exactly without incorporating additional assumptions, even if the mixing process **A** is known [10]. Additional assumptions include knowledge about the geometry, or detailed knowledge about the source distributions [22]. For example, the source signals are assumed to be sparsely distributed - either in the time domain, in the frequency domain or in the time-frequency (T-F) domain [8], [23]–[26]. Sparse sources have a limited overlap in the T-F domain. The validity of non-overlapping sources in the T-F domain comes from the observation that the spectrogram of a mixture is approximately equal to the maximum of the individual spectrograms in the logarithmic domain [27]. When the source signals do not overlap in the time-frequency domain, high-quality reconstruction can be obtained [8]. The property of non-overlapping sources in the T-F domain has been denoted as the W-disjoint orthogonality [28]. Given the short-time Fourier transform (STFT) of two speech signals $S_i(\omega, t)$ and $S_j(\omega, t)$, the W-disjoint orthogonality property can be expressed as

$$S_i(\omega, t)S_j(\omega, t) = 0, \forall i \neq j, \forall \omega, t, \tag{4}$$

where t is the time frame index and ω is the discrete frequency index. This property holds, for example, when tones are disjoint in frequency.

However, there is overlap between the source signals but good separation can still be obtained by applying a binary time-frequency mask to the mixture [24], [8]. In *computational auditory scene analysis* [29], the technique of T-F masking has been commonly used for many years (see e.g. [30]). Here, source separation is based on organizational cues from auditory scene analysis [31]. Binary masking is consistent with perceptual constraints regarding human ability to hear and segregate sounds [32]. Especially, time-frequency masking is closely related to the prominent phenomenon of auditory masking [33]. More recently the technique has also become popular in the ICA community to deal with non-overlapping sources in the T-F domain [28]. T-F masking is applicable to source separation/segregation using one microphone [30], [34]–[36] or more than one microphone [8], [24], [37]. T-F masking is typically applied as a binary mask. For a binary mask, each T-F unit (the signal element at a particular time and frequency) is either weighted by one or by zero. In order to reduce artifacts, soft masks may also be applied [38]. Also by decreasing the downsampling factor in the signal analysis and synthesis, a reduction of musical noise is obtained [39].

An advantage of using a T-F binary mask is that only a binary decision has to be made [32]. Such a decision can be based on clustering from different ways of direction-of-arrival estimation [8], [24], [28], [37], [40]. ICA has been used in different combinations with the binary mask [40]–[42]. In [40], separation is performed by removing $N - M$ signals by masking and then applying ICA in order to separate the remaining M signals. In [41], ICA has been used the other way around. Here, ICA is applied to separate two signals by using two microphones. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio (SNR).

In this paper, we propose a novel approach to separating an arbitrary number of speech signals. Based on the output of a square (2×2) ICA algorithm and binary T-F masks, our approach iteratively segregates signals from a mixture until an estimate of each signal is obtained. Our method is applicable to both instantaneous and convolutive mixtures. A preliminary version of our work has been presented in [43], where we demonstrated the ability of our proposed framework to separate up to six speech mixtures from two instantaneous mixtures. In [44] it has been demonstrated that the approach can be used to segregate stereo music recordings into single instruments or singing voice. In [45] we described an extension to separate convolutive speech mixtures.

The paper is organized as follows. In Section II, we show how instantaneous real-valued ICA can be interpreted geometrically and how the ICA solution can be applied to underdetermined mixtures. In Sections III and IV we develop a novel algorithm that combines ICA and binary T-F masking in order to separate instantaneous as well as convolutive underdetermined speech mixtures. In Section V, we systematically evaluate the proposed method and compare it to existing methods. Further discussion is given in Section VI, and Section VII concludes the paper.

II. Geometrical interpretation of instantaneous ICA

We assume that there is an unknown number of acoustical source signals but only two microphones. It is assumed that each source signal arrives from a distinct direction and no reflections occur, i.e., we assume an anechoic environment in our mixing model. We assume that the source signals are mixed by an instantaneous time-invariant mixing matrix as in Eq. (1). Due to delays between the microphones, instantaneous ICA with a real-valued mixing matrix usually is not applicable to signals recorded at an array of microphones. Nevertheless, if the microphones are placed at the exact same location and have different gains for different directions, the separation of delayed sources can be approximated by the instantaneous mixing model [46]. Hereby, a combination of microphone gains corresponds to a certain directional pattern. The assumption that the microphones are placed at the exact same location can be relaxed. A similar approximation of delayed mixtures to instantaneous mixtures is provided in [47]. There, the differences between closely spaced omnidirectional microphones are used to create directional patterns, where instantaneous ICA can be applied. In the AppendixA, we show how the recordings from two closely spaced omnidirectional microphones can be used to make two directional microphone gains.

Therefore, a realistic assumption is that two directional microphone responses recorded at the same location are available. For evaluation purposes, we have chosen appropriate microphone responses; the frequency independent gain responses are chosen as functions of the direction θ as $r_1(\theta) = 1 + 0.5\cos(\theta)$ and $r_2(\theta) = 1 - 0.5\cos(\theta)$, respectively. The two microphone responses are shown in Fig. 1. Hence, instead of having a mixing system where a given microphone delay

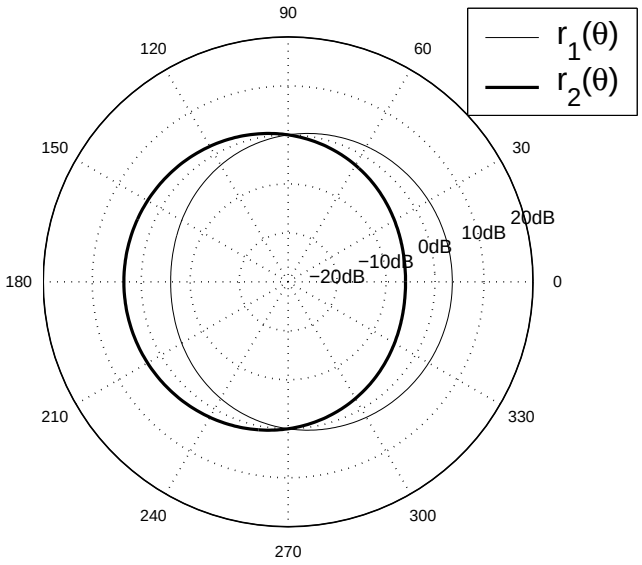


Fig. 1. The two directional microphone responses are shown as function of the direction θ.

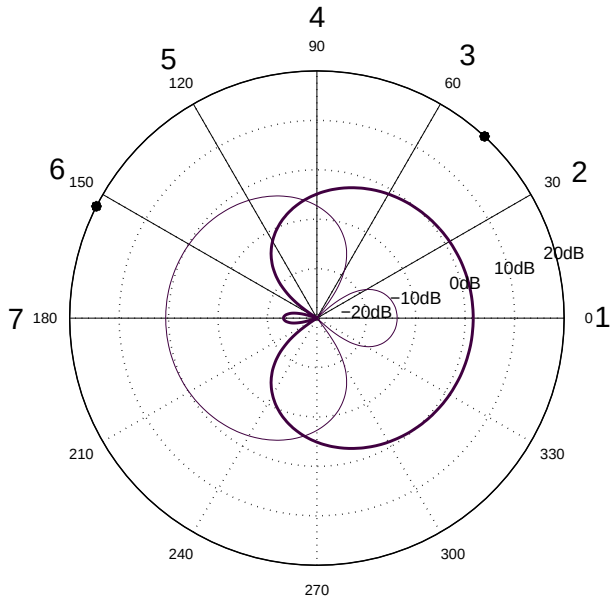


Fig. 2. The polar plots show the gain for different directions. ICA is applied with two sensors and seven sources. The two dots at the outer perimeter show the null directions. We see that each row of the 2×2 ICA solution can make just one null direction in the interval $0^\circ \leq \theta \leq 180^\circ$. Symmetric nulls exist in the interval $180^\circ \leq \theta \leq 360^\circ$. The lines pointing out from the origin denote the true direction of the seven numbered speech sources. The ICA solution tends to place the null towards sources spatially close to each other, and each of the two outputs represents a group of spatially close signals.

corresponds to a given direction, a given set of microphone gains corresponds to a certain direction, and the mixing system is given by

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix}. \quad (5)$$

For the instantaneous case, the separation matrix $\mathbf{W}$ can be regarded as direction-dependent gains. For an $M \times M$ separation matrix, it is possible to have at most $M-1$ null directions, i.e., directions from which the interference signal is canceled out, see e.g. [48], [49]. Signals arriving from other directions are not completely canceled out, and they thus have a gain greater than $-\infty$ dB.

Now consider the case where $N \geq M = 2$. When there are only two mixed signals, a standard ICA algorithm only has two output signals $\mathbf{y}(n) = [y_1(n), y_2(n)]^T$. Since the number of separated signals obtained by (2) is smaller than the number of source signals, $\mathbf{y}$ does not contain the separated signals. Instead, if the noise term is disregarded, $\mathbf{y}$ is another linear superposition of the source signals, i.e.

$$\mathbf{y} = \mathbf{G}\mathbf{s}, \quad (6)$$

where the weights are given by $\mathbf{G} = \mathbf{W}\mathbf{A}$ instead of just $\mathbf{A}$ as in (1). Thus, $\mathbf{G}$ just corresponds to another weighting of each of the source signals depending on θ. These weights make $y_1(n)$ and $y_2(n)$ as independent as possible even though $y_1(n)$ and $y_2(n)$ themselves are not single source signals. This is illustrated in Fig. 2. The figure shows the two estimated spatial responses from $\mathbf{G}(\boldsymbol{\theta})$ in the underdetermined case. The response of the m'th output is given by $\mathbf{g}_m(\theta) = |\mathbf{w}_m^T \mathbf{a}(\theta)|$, where $\mathbf{w}_m$ is the separation vector from the m'th output and $\mathbf{a}(\theta) = [r_1(\theta), r_2(\theta)]^T$ is the mixing vector for the arrival direction θ [48]. By varying θ over all possible directions, directivity patterns can be created as shown in Fig. 2. The estimated null placement is illustrated by the two round dots placed at the perimeter of the outer polar plot. The lines pointing out from the origin illustrate the direction of the seven source signals. Here, the sources are equally distributed in the interval $0^\circ \leq \theta \leq 180^\circ$. As shown in the figure, typically the nulls do not cancel single sources out. Rather, a null is placed at a direction pointing towards a *group* of sources which are spatially close to each other. Here, it can be seen that in the first output, $y_1(n)$, the signals 5, 6 and 7 dominate and in the second output, $y_2(n)$, the signals 1, 2 and 3 dominate. The last signal, 4 exists with almost equal weight in both outputs. As we show in Section III, this new weighting of the signals can be used to estimate binary masks reliably. Similar equivalence has been shown between ICA in the frequency domain and adaptive beamforming [49]. In that case, for each frequency, $\mathbf{Y}(\omega) = \mathbf{G}(\omega)\mathbf{S}(\omega)$.

III. Blind Source Extraction with ICA and Binary Masking

A. Algorithm for instantaneous mixtures

The input to our algorithm is the two mixtures x_1 and x_2 of duration N_s. The algorithm can be divided into three main parts: a *core procedure*, a *separation stage* and a *merging stage*. The three parts are presented in Fig. 3, Fig. 4 and Fig. 5, respectively.

1) Core procedure: Fig. 3 shows the *core procedure*. The core procedure is performed iteratively for a number of cycles in the algorithm. The inputs to the core procedure are two input

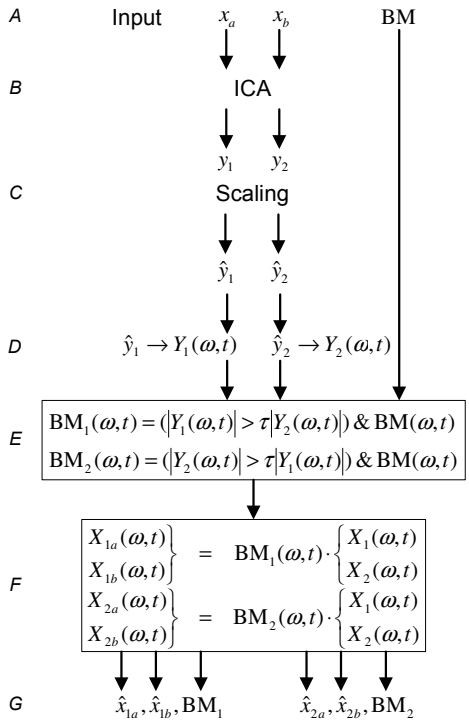


Fig. 3. Flowchart showing the *core procedure* of the algorithm. The algorithm has three input signals: The two input mixtures $x_a = [x_a(0), x_a(1), \cdots, x_a(N_s)]$ and $x_b = [x_b(0), x_b(1), \cdots, x_b(N_s)]$, and a binary mask which has been applied to the two original mixtures in order to obtain x_a and x_b. Source separation by ICA is applied to the two original signals in order to obtain y_1 and y_2. $\hat{y}_1$ and $\hat{y}_2$ are obtained by normalizing the two signals with respect to the variance. The re-scaled signals are transformed into the T-F domain, where the two binary masks are obtained by comparing the corresponding T-F units of the two T-F signals and multiplying by the input binary mask to prevent re-introduction of already masked T-F units. The two estimated masks are then applied in the T-F domain to the original signals $x_1 \rightarrow X_1(\omega,t)$ and $x_2 \rightarrow X_2(\omega,t)$. The output consists of the two estimated binary masks and the four masked signals.

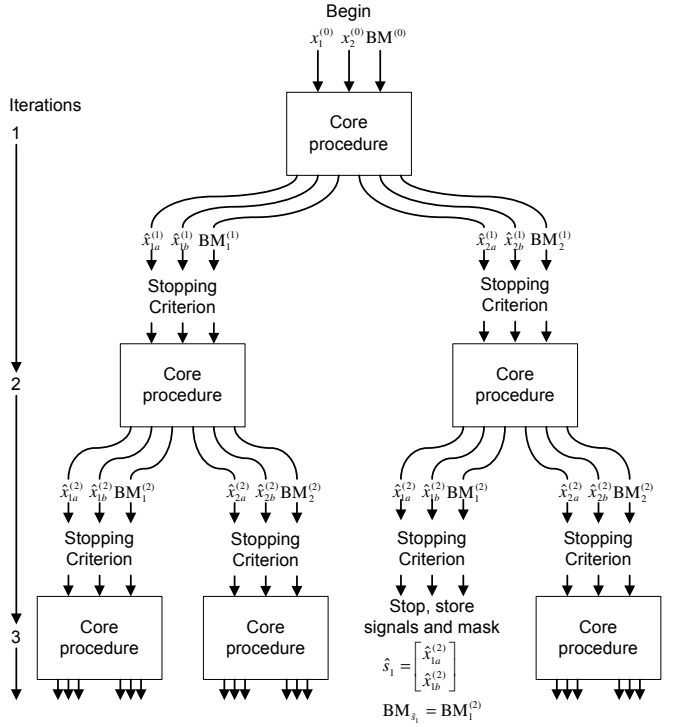


Fig. 4. The *separation stage*. Separation is performed iteratively by the core procedure as described in Fig. 3. The stopping criterion is applied to each set of outputs from the core procedure. If the output consists of more than one speech signal, the core procedure is applied again. If the output consists of only a single source signal, the output and its corresponding mask are stored. The core procedure is applied to the outputs iteratively until all outputs consist of only a single signal. The outputs are stored either as a candidate for a separated stereo sound signal $\hat{s}$ or a separated stereo signal of poor quality $\hat{p}$.

mixtures x_a and x_b and a binary mask (step *A*), which has been applied to the original signals x_1 and x_2 in order to obtain x_a and x_b. In the initial application of the core procedure, $x_a = x_1$ and $x_b = x_2$, and BM is all ones.

As described in the previous section, a two-input two-output ICA algorithm is applied to the input mixtures, regardless of the number of source signals that actually exist in the mixture (step *B*). The two outputs y_1 and y_2 from the ICA algorithm are arbitrarily scaled (step *C*). Since the binary mask is estimated by comparing the amplitudes of the two ICA outputs, it is necessary to solve the scaling problem. In [43], we solved the scaling problem by using the knowledge about the microphone responses. Here we use a more 'blind' method to solve the scaling ambiguity. As proposed in [10], we assume that all source signals have the same variance and the outputs are therefore scaled to have the same variance. The two re-scaled output signals, $\hat{y}_1$ and $\hat{y}_2$ are transformed into the frequency domain (step *D*), e.g. by use of the STFT so that two spectrograms are obtained:

$$\hat{y}_1 \rightarrow Y_1(\omega,t) \quad (7)$$

$$\hat{y}_2 \rightarrow Y_2(\omega,t), \quad (8)$$

where ω denotes the frequency and t the time window index. From the two time-frequency signals, two binary masks are estimated. The binary masks are determined for each T-F unit by comparing the amplitudes of the two spectrograms (step *E*):

$$\mathrm{BM}_1(\omega,t) = \begin{cases} 1, & \text{if } |Y_1(\omega,t)| > \tau|Y_2(\omega,t)|; \\ 0, & \text{otherwise.} \end{cases} \quad \forall \omega, t \quad (9)$$

$$\mathrm{BM}_2(\omega,t) = \begin{cases} 1, & \text{if } |Y_2(\omega,t)| > \tau|Y_1(\omega,t)|; \\ 0, & \text{otherwise.} \end{cases} \quad \forall \omega, t \quad (10)$$

where τ is a parameter. The parameter τ in (9) and (10) controls how sparse the mask should be, i.e., how much of the interfering signals should be removed at each iteration. If $\tau = 1$, the two estimated masks together contain the same number of retained T-F units (i.e. equal to 1) as the previous mask. If $\tau > 1$, the combination of the two estimated masks is more sparse, i.e. having fewer retained units, than the previous binary mask. This is illustrated in Fig. 6. In general, when $\tau > 1$, the convergence is faster at the expense of a sparser resulting mask. When the mask is sparser, musical noise becomes more audible. The performance of the algorithm is considered for $\tau = 1$ and $\tau = 2$. We do not consider the case where $0 < \tau < 1$ as some T-F units would be assigned the value '1' in both estimated masks.

In order to ensure that the binary mask becomes sparser for

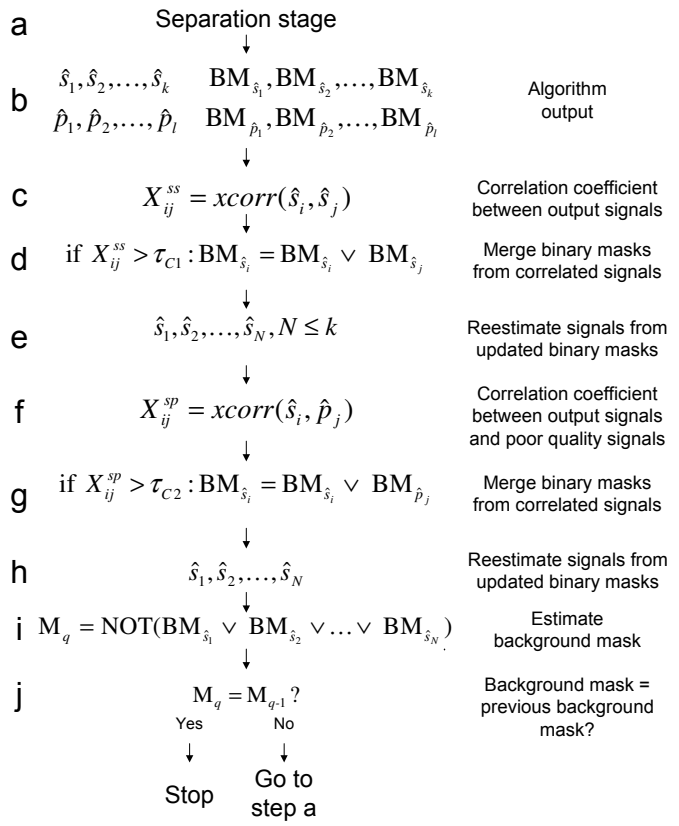


Fig. 5. Flowchart showing the steps of the *merging stage*. The details of the separation stage in step 'a' are shown in Fig. 3 and in Fig. 4. From the separation stage, the outputs shown in step 'b' are available. $\hat{s}_1, \ldots, \hat{s}_k$ denote the k separated signals, and $\hat{p}_1, \ldots, \hat{p}_l$ denotes the l separated signals of poor quality. BM denotes the corresponding binary mask of the estimated signal. The outputs from the main algorithm are further processed in order to improve the separated signals. Masks of output signals which are correlated are merged. Also masks output signals which are correlated with signals of poor quality are merged with these masks. A *background mask* is estimated from T-F units that have not been used so far. This mask is used to execute the main algorithm again. If the background mask has not changed, the segregated signals are not changed any further and the algorithm stops.

every iteration, a simple logical AND operation between the previous mask and the estimated mask is applied.

Next, each of the two binary masks is applied to the original mixtures in the T-F domain (step *F*), and by this non-linear processing, some of the speech signals are *attenuated* by one of the masks while other speakers are attenuated by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT (step *G*).

Time-frequency decomposition can be obtained in many ways, of which the STFT is only one way. The STFT has a linear frequency scale. A linear frequency scale does not accord well with human perception of sounds. The frequency representation in the human ear is closer to a logarithmic scale. The frequency resolution at the low frequencies is much higher than that at the high frequencies [33]. Therefore, T-F decomposition, where the frequency spacing is logarithmic may be a better choice than a linear scale. T-F decomposition based on models of the cochlea are termed *cochleagrams* [29]. Different filterbanks can be used in order to mimic the cochlea, including the Gammatone filterbank [50]. Frequency warping of a spectrogram is another option, e.g. to fit the Bark frequency scale [51].

2) Separation stage: Fig. 4 shows the *separation stage*, i.e. how the core procedure is applied iteratively in order to segregate all the source signals from the mixture. At the beginning, the two recorded mixtures are used as input to the core procedure. The initial binary mask, $\mathrm{BM}^{(0)}$ has the value '1' for all T-F units. A stopping criterion is applied to the two sets of masked output signals. The masked output signals are divided into three categories defined by the stopping criterion in Section IV:

1) The masked signal is of poor quality.
2) The masked signal consists of mainly one source signal.
3) The masked signal consists of more than one source signal.

In the first case, the poor quality signal is stored for later use and marked as a poor quality signal. We denote these signals as $\hat{p}$. When we refer to a signal of poor quality, we mean a signal whose mask only contains few T-F units. Such a signal is distorted with many artifacts. In the second case, the signal is stored as a candidate for a separated source signal. We denote those signals as $\hat{s}$. In the third case, the masked signal consists of more than one source. Further separation is thus necessary, and the core procedure is applied to the signals. T-F units that have been removed by a previous mask cannot be re-introduced in a later mask. Thus, for each iteration, the estimated binary masks become sparser. This iterative procedure is followed until no more signals consist of more than one source signal.

3) Merging stage: The objective of our proposed method is to segregate all the source signals from the mixture. Because a signal may be present in both ICA outputs, there is no guarantee that two different estimated masks do not lead to the same separated source signal. In order to increase the probability that all the sources are segregated and no source has been segregated more than once, a *merging stage* is applied. Further, the merging stage can also improve the quality of the estimated signals. The merging steps are shown in Fig. 5. The output of the separation stage (step a) is shown in step b. The output of the algorithm consists of the k segregated sources, $\hat{s}_1, \ldots, \hat{s}_k$, the l segregated signals of poor quality, $\hat{p}_1, \ldots, \hat{p}_l$, and their corresponding binary masks. In the merging stage, we identify binary masks that mainly contain the same source signal. A simple way to decide whether two masks contain the same signal is to consider the correlation between the masked signals in the time domain. Notice that we cannot find the correlation between the binary masks. The binary masks are disjoint with little correlation. Because we have overlap between consecutive time frames, segregated signals that originate from the same source are correlated in the time domain.

In step c, the correlation coefficients between all the separated signals are found. If the normalized correlation coefficient between two signals is greater than a threshold τ_{C1}, a new signal is created from a new binary mask as shown in step d and e. The new mask is created by applying the logical OR operation to the two masks associated with the two correlated

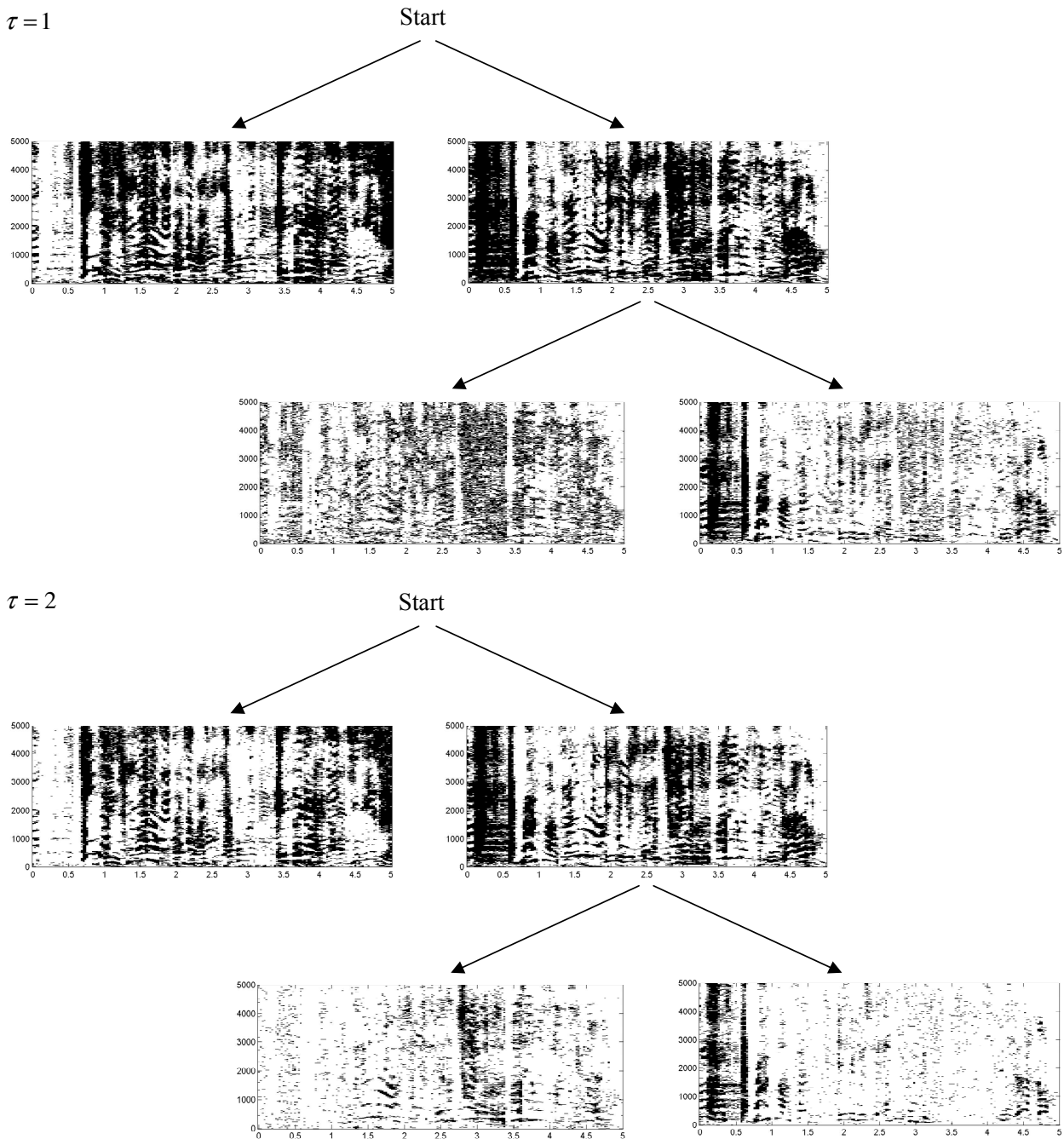


Fig. 6. The first two iterations for the estimations of the binary masks. Black indicates '1', and white '0'. For each iteration, two new masks are estimated by comparison of the ICA output as shown in equations (9) and (10). The previous mask ensures that no T-F units are re-introduced. The plot above shows the case of $\tau = 1$. When $\tau = 1$, the estimated masks contain the same T-F units as the mask in the previous iteration. The plot below shows the case of $\tau = 2$. Here the two estimated masks together contain less T-F units than the binary mask at the previous iteration. Therefore τ can be used to control the convergence speed. The separation performance with the $\tau = 1$ and $\tau = 2$ is presented in Table V and VI, respectively.

signals. Here, we just find the correlation coefficients from one of the two microphone signals and assume that the correlation coefficient from the other channel is similar.

Even though a segregated signal is of poor quality, it might still contribute to improve the quality of the extracted signals. Thus, the correlation between the signals with low quality (energy) and the signals that contain only one source signal is found (step f). If the correlation is greater than a threshold τ_{C2}, the mask of the segregated signal is expanded by merging the mask of the signal of poor quality (step g and h). Hereby the overall quality of the new mask should be higher, because the new mask is less sparse. After the correlations between the output signals have been found, some T-F units still have not been assigned to any of the source signal estimates. As illustrated in Fig. 7, there is a possibility that some of the sources in the mixture have not been segregated. In the

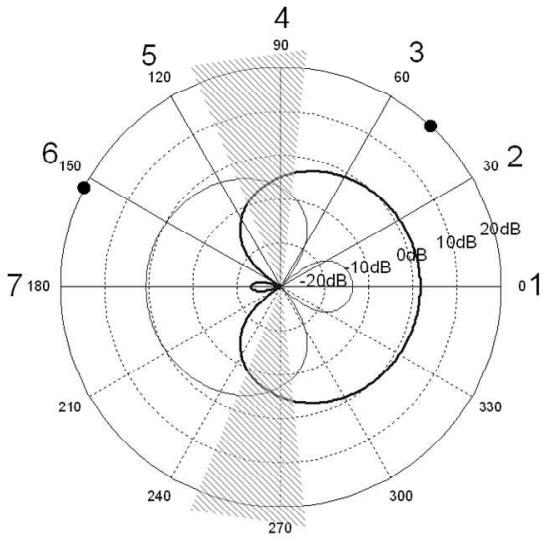


Fig. 7. As in Fig. 2 the polar plots show the gain for different directions. Comparison between the gains determines the binary masks. Within the shaded areas, the gain is almost equal. Source signals that arrive from a direction close to where the gains are almost equal will (depending on the parameter τ) either exist in both masked signals or in none of the masked signals. Therefore, the algorithm may fail to segregate such source signals from the mixture.

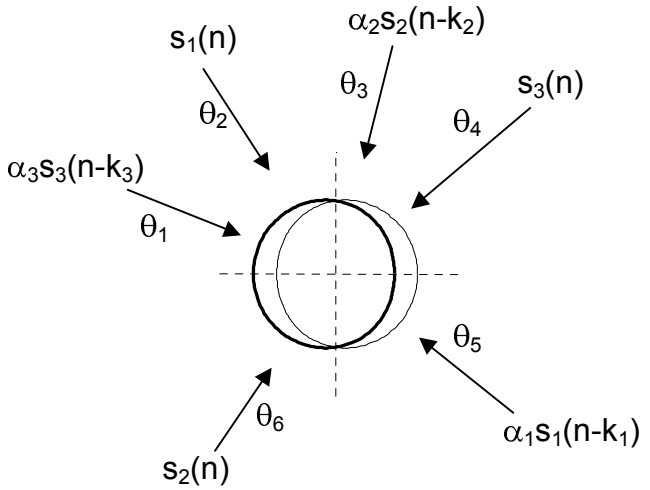


Fig. 8. A simple reverberant environment with three sources each having one reflection. As in Fig. 1 the impinging signals are recorded by a two-microphone array with directional responses, so that each direction corresponds to a certain set of directional microphone responses. Here, each reflection can be regarded as a single source impinging the microphone array.

direction where the gains from the two ICA outputs are almost equal, there is a higher uncertainty on the binary decision, which means that a source in that area may appear in both outputs. Furthermore, if $\tau > 1$ some T-F units in the shaded area of Fig. 7 are assigned the value '0' in both binary masks. Therefore, sources are assumed to exist in the T-F units which have not been assigned to a particular source yet. Thus, a *background mask* is created from all the T-F units which have not been assigned to a source (step i). The background mask is then applied to the original two mixtures, and possible sounds that remain in the background mask are hereby extracted. The separation algorithm is then applied to the remaining signal to ensure that there is no further signal to extract. This process continues until the remaining mask does not change any more (step j). Notice that the final output signals are maintained as two signals.

B. Modified algorithm for convolutive mixtures

In a reverberant environment, reflections from the signals generally arrive from different directions. In this situation, the mixing model is given by (3). Again, we assume that the sounds are recorded by a two-microphone array with directional responses given in Fig. 1. A simple reverberant environment is illustrated in Fig. 8. Here three sources $s_1(n)$, $s_2(n)$ and $s_3(n)$ are impinging the two-microphone array and direction-dependent gains are obtained. Also one reflection from each of the sources is recorded by the directional microphones: $\alpha_1 s_1(n-k_1)$, $\alpha_2 s_2(n-k_2)$ and $\alpha_3 s_3(n-k_3)$. In this environment, we can write the mixture with an instantaneous mixing model $\mathbf{x} = \mathbf{As}$ with $\mathbf{s} = [\alpha_3 s_3(n-k_3), s_1(n), \alpha_2 s_2(n-k_2), s_3(n), \alpha_1 s_1(n-k_1), s_2(n)]^T$ and

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_6) \\ r_2(\theta_1) & \cdots & r_2(\theta_6) \end{bmatrix}. \tag{11}$$

We can therefore apply the iterative instantaneous ICA algorithm to the mixture, and we can segregate the convolutive mixture into numerous components, as independent as possible, where each component is a source or a reflection impinging from a certain direction. Similarly, a merging stage can determine if two segregated components originate from the same source.

When the method is applied to reverberant mixtures, we observe that the estimated binary masks becomes more frequency dependent so that the binary mask for some frequencies mainly contains zeroes and for other frequency bands mainly contains ones. This results in band-pass filtered versions of the segregated signals. For example, one binary mask mainly contains the high-frequency part of a speech signal, while another mask mainly contains a low-frequency part of the same speech signal. This high-pass and low-pass filtered versions are poorly correlated in the time-domain. In order to merge these band-pass filtered speech signals that originate from the same source, we compute the correlation between the envelopes of the signals instead. This approach has successfully been applied in frequency domain ICA in order to align permuted frequencies [52], [53]. The following example shows that the envelope correlation is a better merging criterion than just finding the correlation between the signals, when the signals are bandpass-filtered.

Two speech signals A and B with a sampling rate of 10 kHz are each convolved with a room impulse response having $T_{60} = 400$ ms. Both signals are divided into a high-frequency (HF) part, and a low frequency (LF) part. Hereby four signals A_{LF}, A_{HF}, B_{LF}, and B_{HF} are obtained. The two LF signals are obtained from binary masks which contain ones

for frequencies below 2500 Hz and zeros otherwise, and the two HF signals are obtained from binary masks which contain ones for frequencies above 2500 Hz and zeros otherwise. We now find the correlation coefficients between the four signals and the envelopes. The envelope can be obtained in different ways. The envelope $\mathcal{E}$ of the signal $x(n)$ can be calculated as [54]

$$\mathcal{E}(x(n)) = |x(n) + j\mathcal{H}(x(n))|, \tag{12}$$

where $\mathcal{H}(x(t))$ denotes the Hilbert transform, and j denotes the imaginary unit. Alternatively, we can obtain a smoother estimate $\hat{\mathcal{E}}$ as

$$\hat{\mathcal{E}}(x(n)) = \hat{\mathcal{E}}(x(n-1)) + \alpha(n)\big(|x(n)| - \hat{\mathcal{E}}(x(n-1))\big), \tag{13}$$

where

$$\alpha = \begin{cases} 0.04, & \text{if } |x(n)| - \hat{\mathcal{E}}(x(n-1)) > 0; \\ 0.01, & \text{if } |x(n)| - \hat{\mathcal{E}}(x(n-1)) < 0. \end{cases} \tag{14}$$

The above values of α have been found experimentally. The attack time and release time of the low-pass filter have been chosen differently in order to track the onsets easily. We initialize (13) by setting $\hat{\mathcal{E}}(x(0)) = 0$.

To prevent the DC component of the envelope from contributing to the correlation, the DC components are removed from the envelopes by a high-pass filter, before the correlation coefficient between the envelopes is computed. In Table I, the correlation coefficients between the four signals have been found, as well as the correlations between the envelopes and the smoothed envelopes. It is desirable that the correlation between signals that originate from the same source be high while the correlation between different signals be low. As it can be seen, the correlation coefficients between the signals do not indicate that A_{LF} and A_{HF} (or B_{LF} and B_{HF}) belong to the same source signal. When the correlation coefficients between the envelopes are considered, the correlations between A_{LF} and A_{HF} (or B_{LF} and B_{HF}) are a little higher than the cross-correlation between the source signals. The best result is obtained for the correlation between the smoothed envelopes. Here the correlations between A_{LF} and A_{HF} (or B_{LF} and B_{HF}) are significantly higher than the correlations between the different sources. In the reverberant case, we thus merge masks based on correlation between the smoothed envelope. We have also tried to apply the envelope-based merging criterion in the instantaneous case, but found that the simple correlation-based criterion gives better results. The reason, we suspect, is that the temporal fine structure of a signal that is present in the instantaneous case but weakened by reverberation is more effective than the signal envelope for revealing correlation.

TABLE I
CORRELATION BETWEEN HIGH- AND LOW-PASS FILTERED SPEECH SIGNALS, THE ENVELOPE OF THE SIGNALS AND THE SMOOTHED ENVELOPE OF THE SIGNALS.

	A_{LF}	A_{HF}	B_{LF}	B_{HF}
A_{LF}	1	0.0006	0.0185	0.0001
A_{HF}		1	0.0001	0.0203
B_{LF}			1	0.0006
B_{HF}				1

	$\mathcal{E}(A_{\text{LF}})$	$\mathcal{E}(A_{\text{HF}})$	$\mathcal{E}(B_{\text{LF}})$	$\mathcal{E}(B_{\text{HF}})$
$\mathcal{E}(A_{\text{LF}})$	1	0.0176	0.0118	0.0131
$\mathcal{E}(A_{\text{HF}})$		1	0.0106	0.0202
$\mathcal{E}(B_{\text{LF}})$			1	0.0406
$\mathcal{E}(B_{\text{HF}})$				1

	$\hat{\mathcal{E}}(A_{\text{LF}})$	$\hat{\mathcal{E}}(A_{\text{HF}})$	$\hat{\mathcal{E}}(B_{\text{LF}})$	$\hat{\mathcal{E}}(B_{\text{HF}})$
$\hat{\mathcal{E}}(A_{\text{LF}})$	1	0.0844	0.0286	0.0137
$\hat{\mathcal{E}}(A_{\text{HF}})$		1	0.0202	0.0223
$\hat{\mathcal{E}}(B_{\text{LF}})$			1	0.0892
$\hat{\mathcal{E}}(B_{\text{HF}})$				1

IV. STOPPING CRITERION

As already mentioned, it is important to decide whether the algorithm should stop or the processing should repeat. The algorithm should stop when the signal consists of only one source or when the mask is too sparse (hence the quality of the resulting signal will be poor). Otherwise, the separation procedure should continue. When there is only one source in the mixture, the signal is expected to arrive only from one direction and thus the rank of the mixing matrix is one. We propose a stopping criterion based on the covariance matrix of the masked sensor signals. An estimate of the covariance matrix is found as

$$\mathbf{R}_{xx} = \langle \mathbf{x}\mathbf{x}^T \rangle = \frac{1}{N_s}\mathbf{x}\mathbf{x}^T, \tag{15}$$

where N_s is the number of samples in $\mathbf{x}$. By inserting (1), and assuming that the noise is independent with variance σ^2, the covariance can be written as function of the mixing matrix and the source signals:

$$\begin{aligned} \mathbf{R}_{xx} &= \langle (\mathbf{A}\mathbf{s} + \boldsymbol{\nu})(\mathbf{A}\mathbf{s} + \boldsymbol{\nu})^T \rangle & (16) \\ &= \mathbf{A}\langle \mathbf{s}\mathbf{s}^T \rangle \mathbf{A}^T + \langle \boldsymbol{\nu}\boldsymbol{\nu}^T \rangle & (17) \\ &= \mathbf{A}\langle \mathbf{s}\mathbf{s}^T \rangle \mathbf{A}^T + \sigma^2 \mathbf{I} & (18) \\ &= \boldsymbol{\Psi} + \sigma^2 \mathbf{I}, & (19) \end{aligned}$$

where $\boldsymbol{\Psi} = \mathbf{A}\mathbf{R}_{ss}\mathbf{A}^T$ of size $M \times M$. We assume that the masked sensor signal consists of a single source if the condition number (based on the 2-norm) [55] is greater than a threshold τ_c, i.e.

$$\text{cond}(\mathbf{R}_{xx}) > \tau_c. \tag{20}$$

A high condition number indicates that the matrix is close to being singular. Since $\mathbf{R}_{xx}$ is symmetric and positive definite, $\text{cond}(\mathbf{R}_{xx}) = \max \text{eig}(\mathbf{R}_{xx}) / \min \text{eig}(\mathbf{R}_{xx})$, where $\text{eig}(\mathbf{R}_{xx})$ is the vector of eigenvalues of $\mathbf{R}_{xx}$. Because the desired signals are speech signals, we bandpass filter the masked mixed signals before we calculate the covariance matrix, so that only frequencies where speech dominates are considered. The cutoff frequencies of the bandpass filter are chosen to be 500 and 3500 Hz.

In order to discriminate between zero and one source signal, we consider the power of the masked signal. If the power of the masked signal has decreased by a certain amount compared to the power of the original mixture, the signal is considered to be of poor quality. We define this amount by the parameter τ_E, which is measured in dB.

This stopping criterion is applied for instantaneous as well as convolutive mixtures. In the case of convolutive mixtures, the stopping criterion aims at stopping when the energy of

the segregated signal mainly comes from a single direction, i.e. the iterative procedure should stop when only a single reflection from a source remains in the mixture. Note that, as illustrated in Fig. 8, our algorithm for convolutive mixtures treats each reflection as a distinct sound source. Because many reflections have low energy compared to the direct path, a high number of segregated signals of poor quality are expected in the reverberant case.

V. Evaluation

A. Evaluation Metrics

When using a binary mask, it is not possible to reconstruct the speech signal perfectly, because the signals partly overlap. An evaluation method that takes this into account is therefore used [56]. As a computational goal for source separation, the *ideal binary mask* has been suggested [32]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the signal to the energy of all the interfering signals. Whenever the signal energy is higher within a T-F unit, the T-F unit is assigned the value '1' and whenever the combined interfering signals have more energy, the T-F unit is assigned the value '0'. The ideal binary mask produces the optimal SNR gain of all binary masks in terms of comparing with the entire signal [34].

As in [34], for each of the separated signals, the percentage of energy loss P_{EL} and the percentage of noise residue P_{NR} are calculated:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \quad (21)$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \quad (22)$$

where $O(n)$ is the estimated signal, and $I(n)$ is the signal re-synthesized after applying the ideal binary mask. $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. The performance measure P_{EL} can be regarded as a weighted sum of the T-F unit power present in the ideal binary mask, but absent in the estimated mask, while the performance measure P_{NR} can be regarded as a weighted sum of the T-F unit power present in the estimated binary mask, but absent in the ideal binary mask.

Also the output signal-to-noise ratio (SNR_o) can be measured. Here the SNR_o is defined using the re-synthesized speech from the ideal binary mask as the ground truth

$$\text{SNR}_\text{o} = 10\log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n)-O(n))^2}\right]. \quad (23)$$

If instead the original signal is used as the ground truth in the numerator in (23), the relatively low target energy from the T-F units that have been assigned the value '0' will also contribute. Because there is good perceptual correlation between the true speech signal and the resynthesized speech signal from the ideal mask [32], we should not let the inaudible values of the true signal contribute disproportionately to the SNR estimation. Therefore, it is better to use the ideal mask as the ground truth. Also the signal-to-noise ratio before separation, the input SNR (SNR_i), is calculated. The SNR_i is the ratio between the desired signal and the interfering signals in the recorded masked mixtures. The SNR gain is measured in dB by

$$\Delta\text{SNR} = \text{SNR}_\text{o} - \text{SNR}_\text{i}. \quad (24)$$

If we instead were using the original signals as ground truth, the SNR gain would be about 1-2 dB lower (see also [34]).

B. Setup and parameter choice

For evaluation, twelve different speech signals - six male and six female - from eleven different languages have been used. All speakers raised voice as if they were speaking in a noisy environment. The duration of each of the signals is five seconds and the sampling frequency is $f_s = 10$ kHz. All the source signals have approximately the same loudness. Separation examples and Matlab source code are available online [57], [58]. The signal positions are chosen to be seven positions equally spaced in the interval $0^\circ \leq \theta \leq 180^\circ$ as shown in Fig. 2. Hereby, the minimum angle between two signals is 30°. During the experiments, each mixture is chosen randomly and each source is randomly assigned to one of the seven positions.

We have experimented with several different random mixtures. Sometimes the method fails in separating all the mixtures. In those cases, typically two segregated signals are merged because they are too correlated, resulting in $N-1$ segregated signals, where one of the segregated signals consists of two source signals which are spatially close to each other. Alternatively, one source signal may occur twice resulting in $N+1$ separated signals. Therefore, as another success criterion we also count the number of times where all N sources in the mixture have been segregated into exactly N signals and each of the N sources are dominating in exactly one of the segregated signals. We call the ratio "correctness of detected source number" or "Correct #" in the result tables. We then calculate the average performance from those where the number of sources has been correctly detected when the algorithm stops. Although not all signals are correctly separated, it is still useful for some applications to recover some of the signals. Subjective listening could determine which of the source signals in the mixture the segregated signal is closest to. Here we use an automatic method to determine the pairing between the segregated signal and a source signal by comparing the corresponding estimated mask of the segregated signal and the ideal masks of different source signals. The source signal whose corresponding ideal mask is closest (in terms of most number of ones in common) to the estimated mask is determined to correspond to the segregated source. This method correlates well with subjective listening.

Different instantaneous ICA algorithms can be applied to the method. For evaluation we use an implementation of the IN-

TABLE II
ROBUSTNESS OF τ_C AND τ_E FOR INSTANTANEOUS MIXTURES OF $N = 4$ AND $N = 6$ SIGNALS.

	τ_C					
	$N = 4$			$N = 6$		
τ_E	2000	3000	4000	2000	3000	4000
	15.45	15.34	15.24	13.85	14.04	13.87
15	10/10	10/10	9/10	6/10	5/10	5/10
	15.34	15.23	15.18	13.91	13.94	14.06
20	10/10	10/10	10/10	8/10	9/10	6/10
	15.64	15.19	14.36	14.39	13.86	14.06
25	4/10	4/10	5/10	1/10	4/10	6/10

ΔSNR and the number of times
(out of the ten cases) where all signals have been segregated

FOMAX ICA algorithm [13] which uses the BFGS (Broyden-Fletcher-Goldfarb-Shanno) optimization method [59], [60]. Unless otherwise stated, the parameter τ in Equations (9) and (10) is set to $\tau = 1$.

1) Choice of thresholds: Different thresholds have to be chosen. The thresholds have been determined from initial experiments as described below.

Regarding the two correlation thresholds, τ_{C1} and τ_{C2} shown in Fig. 5, our experiments show that most correlations between the time signals are very close to zero. Two candidates for separated signals are merged if the correlation coefficient is greater than 0.1. If τ_{C1} is increased, some signals may not be merged even though they mainly contain the same source. If τ_{C2} is decreased, the probability of merging different source signals is increased. The low energy signals are even less correlated with the candidates for separated signals. Therefore, we have chosen $\tau_{C2} = 0.03$. If τ_{C2} is increased, the masks become sparser, and more artifacts occur. If τ_{C2} becomes smaller, noise from other sources becomes more audible.

The thresholds in the stopping criterion are estimated from the initial experiments too. The condition number related threshold is chosen to be $\tau_C = 3000$. The signal is considered to contain too little energy when the energy of the segregated signal has decreased to $\tau_E = -20$ dB, when the power of a recorded mixture is normalized to 0 dB.

The robustness of the two thresholds τ_C and τ_E has been evaluated. τ_C has been evaluated for the values 2000, 3000 and 4000. Likewise, τ_E has been evaluated for the values 15, 20 and 25 dB. For each pair of τ_C and τ_E ten different random speech mixtures drawn from the pool of twelve speech signals are segregated. The experiment has been performed for mixtures consisting of four or six speech signals. In each case, ΔSNR is measured. Also the number of times (out of ten) where exactly all the sources in the mixture are been segregated is found. The results are reported in Table II. As it can be seen, the ΔSNR does not vary much as function of the two thresholds. The number of times where the method fails to segregate exactly N speech signals from the mixture is minimized for $\tau_C = 3000$ and $\tau_E = 20$ dB, which will be used in the evaluation.

The algorithm could be applied to a mixture several times, each time with different thresholds. Such a procedure could increase the chance of extracting all the sources from the mixture.

TABLE III
PERFORMANCE FOR DIFFERENT WINDOW LENGTHS

Window length	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Correct #
256 (25.6 ms)	9.17	11.38	13.56	44/50
512 (51.2 ms)	6.07	8.62	15.23	46/50
1024 (102.4 ms)	6.86	9.92	14.75	46/50

The window length is given in samples and in milliseconds.
The DFT length is four times the window length.
The number of signals in each instantaneous mixture is $N = 4$.

TABLE IV
COMPARISON BETWEEN JADE AND INFOMAX ICA ALGORITHMS.

Algorithm	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Correct #
JADE	6.20	8.86	15.17	46/50
INFOMAX	6.07	8.62	15.23	46/50

Instantaneous mixtures consisting of four sources have been used.

2) Window function: In [8], the Hamming window is found to perform slightly better than other window functions. In the following, the Hamming window will be used.

3) Window length: Different window lengths have been tried. The overlap factor is selected to be 75%. An overlap factor of 50% has also been considered, but better performance is obtained with 75% overlap.

With an overlap of 75% the separation has been evaluated for window lengths of 256, 512 and 1024 samples, which with $f_s = 10$ kHz give window shifts of 12.8, 25.6 and 51.2 ms, respectively. For a Hamming window the 3 dB bandwidth of the main lobe is 1.30 samples [61]. The frequency (spectral) resolution is thus 50.8, 25.4 and 12.7 Hz, respectively. The DFT length is four times the window length. Hence, the spectrogram resolution is 513, 1025 and 2049, respectively. By selecting a DFT length longer than the window length, the spectrogram becomes smoother, and when listening to the segregated signals, the quality becomes much better too. When the DFT size is longer than the window size, there is more overlap between the different frequency bands. Furthermore, artifacts from aliasing are reduced by zero-padding the window function.

The results are shown in Table III. The average performance is given for fifty random mixtures, each consisting of four speech sources. The highest SNR improvement is achieved for a window length of 512. A similar performance is achieved for the window length of 1024, while the window length of 256 performs a little worse. In the following experiments, we use a window length of 512.

4) ICA algorithm: We have chosen to use the INFOMAX algorithm [13] for evaluation, but other ICA algorithms could be used also. To examine how much the performance of our method depends on the chosen ICA algorithm, we have compared the INFOMAX and the JADE algorithm [62] in the ICA step. In both cases, the code is available online [59], [63]. The two algorithms have been applied to the same fifty mixtures each consisting of four signals drawn from the pool of twelve signals. The results are given in Table IV. As it can be seen, the performance of our method does not depend much on whether the chosen ICA algorithm is the INFOMAX or the JADE algorithm.

TABLE V
EVALUATION WITH RANDOM INSTANTANEOUS MIXTURES CONSISTING OF N SIGNALS.

N	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
2	1.01	2.00	0	18.92	18.92	47/50
3	2.99	4.86	-3.95	12.50	16.45	46/50
4	6.07	8.62	-5.98	9.26	15.23	46/50
5	10.73	13.02	-7.40	5.56	14.27	44/50
6	14.31	13.63	-8.39	5.25	13.64	44/50
7	18.34	22.43	-9.24	4.24	13.48	41/50

The parameter $\tau = 1$.

TABLE VI
EVALUATION WITH RANDOM INSTANTANEOUS MIXTURES CONSISTING OF N SIGNALS.

N	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
2	3.43	0.50	0	18.22	18.22	50/50
3	7.36	2.60	-3.96	11.10	15.06	46/50
4	12.26	4.17	-5.89	8.81	14.70	42/50
5	19.81	6.21	-7.32	6.59	13.91	40/50
6	25.91	8.81	-8.36	5.31	13.67	23/50
7	30.52	11.86	-9.12	3.00	13.46	4/50

The parameter $\tau = 2$.

C. Separation results for instantaneous mixtures

Tables V and VI show the average separation performance for mixtures of N signals for $\tau = 1$ and $\tau = 2$. For each N, the algorithm has been applied fifty times to different speaker mixtures from the pool of twelve speakers at N of the seven random positions.

As it can be seen, the proposed algorithm is capable of separating at least up to seven source signals. It can also be seen that the probability of recovering all N speech signals decreases as N increases. Also, the quality of the separated signals deteriorates when N increases. When N increases, the T-F domain becomes less sparse because of higher overlap between the source signals. When the performance for $\tau = 1$ in Table V is compared with that for $\tau = 2$ in Table VI, it can be seen that the performance is better for $\tau = 1$. However the algorithm with $\tau = 1$ uses more computation time compared to $\tau = 2$. As it can be seen in Table V, the algorithm fails to separate two sources from each other in three cases. This is probably because the masks at some point are merged due to a wrong decision by the merging criterion. In Fig. 9, the ideal binary masks for a source from an example mixture of three speech signals are shown, along with the estimated mask is shown. As it can be seen, the estimated mask is very similar to the ideal masks.

1) Stationarity assumption: The duration of the mixture is important for separation. It is required that the source signals remain at their positions while the data is recorded. Otherwise the mixing matrix will vary with time. Therefore, there is a tradeoff between the number of available samples and the time duration during which the mixing matrix can be assumed to be stationary. Mixtures containing four speech signals have been separated. The duration T is varied between 1 and 5 seconds. The average performance has been found from fifty different mixtures. Since the speech mixtures are randomly picked, one second is selected as the lower limit to ensure that

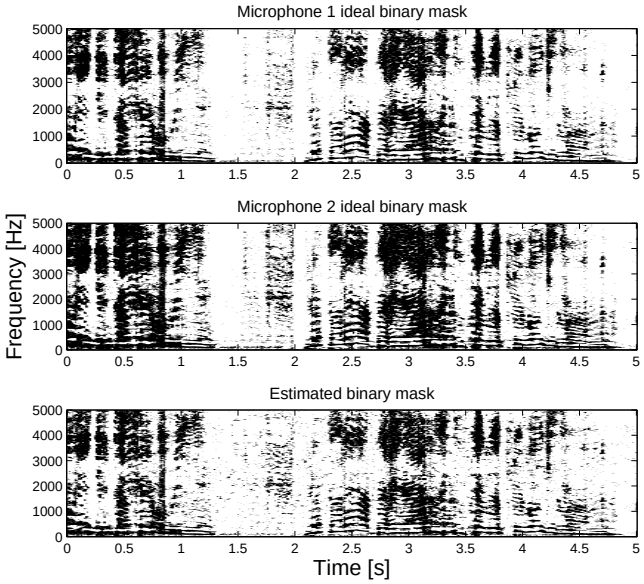


Fig. 9. Separation example. A segregated speech signal from a mixture of three speech signals. The two upper masks show the ideal binary mask for each of the two directional microphones. For this estimated signal, $P_{EL} = 1.38\%$, $P_{NR} = 0.46\%$, and $\Delta\text{SNR} = 20.98$ dB. Notice, unless the ideal masks from both microphones are exactly the same, P_{EL} and P_{NR} are always greater than zero. Perceptually, the segregated signal sounds clean without any artifacts. The separation quality is similar for the two other signals from the mixture.

TABLE VII
EVALUATION OF SEPARATION PERFORMANCE AS FUNCTION OF THE SIGNAL LENGTH T.

T	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
1	7.53	8.83	-6.38	9.44	15.83	34/50
2	7.85	8.23	-5.98	9.00	14.88	43/50
3	6.87	9.69	-6.04	8.80	14.85	46/50
4	7.57	9.05	-6.04	8.81	14.86	46/50
5	6.07	8.62	-5.98	9.26	15.23	46/50

Instantaneous mixtures consisting of four sources have been used.

all four speech signals are active in the selected time frame. The separation results are shown in Table VII. Fifty mixtures of four source signals have been separated and the average performance is shown. As it can be seen, the probability of recovering all the source signals decreases when less data is available. On the other hand, the performance does not increase further for data lengths above three seconds. By listening to the separated signals, we find that among the mixtures where all sources have been successfully recovered, there is no significant difference in the quality of the separated signals.

2) Different loudness levels: In the previous simulations, all the speech signals are approximately equally strong. Now we test the separation performance in situations where the signals in the mixture have different levels of loudness. The mixtures consist of four speech signals, drawn from the pool of twelve signals. Before mixing, the first speech signal is multiplied by 1, the second speech signal is multiplied by 0.5, and the remaining two speech sources are multiplied by 0.25. The average performance from fifty simulations is found. The

TABLE VIII
EVALUATION OF SEPARATION PERFORMANCE AS FUNCTION OF ADDITIVE MICROPHONE NOISE.

Noise	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
-10 dB	15.29	15.52	-6.51	5.95	12.43	19/50
-20 dB	7.42	10.26	-6.02	8.37	14.39	45/50
-30 dB	6.24	8.53	-5.99	9.27	15.26	46/50
-40 dB	6.23	8.72	-5.97	9.19	15.16	47/50
-50 dB	6.39	8.15	-5.98	9.29	15.27	45/50
-60 dB	6.04	8.62	-5.98	9.27	15.25	46/50

Instantaneous mixtures consisting of four sources have been used.

TABLE IX
EVALUATION OF DIRECTIONAL MICROPHONE APPROXIMATION.

Mic. dist.	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Correct #
$d = 1$ cm	7.63	8.84	14.83	17/50
Ideal case	6.07	8.62	15.23	46/50

Anechoic mixtures consisting of four sources have been used.

two strongest sources are segregated in all the examples. In 25 of the 50 simulations, all of the four signals are segregated. On average ΔSNR is 16.57 dB, $P_{EL} = 6.65\%$ and $P_{NR} = 14.64\%$. When we compare to the more difficult case in Table V where all four speakers have equal loudness, we see that the average ΔSNR here is 1 dB better.

3) Microphone noise: In the previous simulations, noise is omitted. We now add white noise to the directional microphone signals with different noise levels. The simulation results are given in Table VIII. The noise level is calculated with respect to the level of the mixtures at the microphone. The mixtures without noise are normalized to 0 dB. As it can be seen from the table, noise levels of up to -20 dB can be well tolerated.

D. Separation results for anechoic mixtures

As mentioned in Section II, directional microphone gains can be obtained from two closely-spaced microphones. Signals impinging at a two-microphone array have been simulated and the directional microphone gains have been obtained as described in the AppendixA. The distance between the microphones is chosen as $d = 1$ cm. Hereby an instantaneous mixture is approximated from delayed sources. With this setup, fifty mixtures each consisting of four speech signals drawn from the pool of twelve speakers have been evaluated. The results are given in Table IX. Because the microphone gain is slightly frequency-dependent, the performance deteriorates compared to the ideal case where the gain is frequency independent, especially for the frequencies above 4 kHz. This is illustrated in Fig. 10. This might be explained by the fact that the approximation $kd \ll 1$ (described in the Appendix) does not hold for higher frequencies. Fortunately, for the perception of speech, the higher frequencies are less important. It can also be seen that the number of times where the exactly four sources have been segregated is decreased. In many cases one source is segregated more than once, which is not merged in the merging stage because the correlation coefficient is too low.

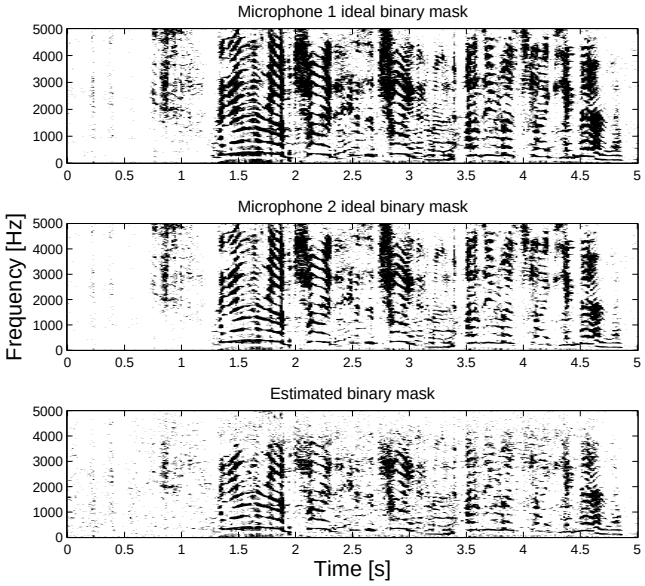


Fig. 10. Separation example. A segregated speech signal from a mixture of four speech signals. The speech signal impinges on an array consisting of two omnidirectional microphones spaced 1 cm apart. The two upper masks show the ideal binary masks for each of the two omnidirectional microphones. Because the directional gains are slightly frequency dependent, the performance for the high frequencies is deteriorated compared to the ideal case when the microphone gain is not frequency dependent, as shown in Fig. 9.

E. Separation results for reverberant recordings

As described in Section III, the method can be applied to recordings of reverberant mixtures. We use recordings from a hearing aid with two closely-spaced, vertically placed omnidirectional microphones. The hearing aid is placed in the right ear of a Head and Torso Simulator (HATS) [64]. Room impulse responses are estimated from different loudspeaker positions. The source signals were then created by convolving the room impulses with the clean speech signals from the pool of twelve speakers.

The impulse responses are found in a reverberant room where the room reverberation time was $T_{60} = 400$ ms. Here reflections from the HATS and the room exist. The microphone distance is 12 mm. The room dimensions were $5.2\times7.9\times3.5$ m and the distance between the microphones and the loudspeakers were 2 m. Impulse responses from loudspeaker positions of 0°, 90°, 135°, and 180° are used. The configuration is shown in Figure 11. Fifty different mixtures consisting of four speakers from the pool of twelve speakers are created. The parameters of the algorithm have to be changed. When reverberation exists, the condition number never becomes as high as the chosen threshold of $\tau_C = 2000$. Therefore we need much lower thresholds. The separation performance is found for different values of τ_C. The remaining thresholds are set to $\tau_E = 25$, $\tau_{C1} = 0.1$ and $\tau_{C2} = 0.05$, with parameter $\tau = 1$. The separation results are provided in Table X. Four sources are not always segregated from a mixture. Therefore we count how many times the algorithm manages to segregate 0, 1, 2, 3 or all four sources from the mixture. This is denoted as

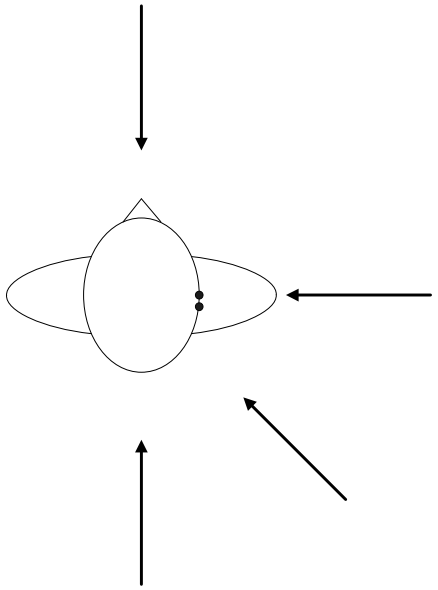


Fig. 11. Room configuration. The Head and Torso Simulator (seen from above) is placed in the middle of a room with a reverberation time of 400 ms. The two-microphone array is placed at the right ear. The distance between the microphones is 12 mm. The four sources arrive from positions of 0°, 90°, 135°, and 180°. The distance from the center of the head to each of the loudspeakers was 2 m. The room dimensions were $5.2 \times 7.9 \times 3.5$ m.

'freq.' in the table. We find the average P_{EL}, P_{NR} and ΔSNR for all these cases. It can be seen that often three of the four signals are segregated from the mixture. The average ΔSNR is around 6 dB. Even though the separation is not as good as in anechoic cases, it is worth noting that instantaneous ICA in the time domain may be used to segregate convolutive mixtures.

Another option is to apply a convolutive ICA algorithm [19] instead of an instantaneous ICA method. This was done in [45]. The advantage of using a convolutive algorithm compared to a instantaneous algorithm is that the convolutive algorithm is able to segregate sources, with larger microphone distances. Still, we have to assume that the convolutive algorithm at each step is able to segregate the sources into two groups, where some sources dominate in one group and other sources dominate in the other group. The stopping criterion from Section IV which is used to discriminate between one and more-than-one signal performs worse under the reverberant condition. Even though the criterion is applied to narrow frequency bands, the performance becomes worse as reported in [65]. In [45], we used a single-microphone criterion based on the properties of speech. There are some advantages of applying an instantaneous ICA as opposed to applying a convolutive ICA algorithm. The instantaneous algorithm is computationally less expensive. Further, frequency permutations which exist in many convolutive algorithms [19] are avoided.

The method used here cannot directly be compared to the method used in [45] which was applied with a much larger microphone distance. In [45], artificial room impulse responses were used with $T_{60} = 160$ ms, and here we have used recorded room impulses with $T_{60} = 400$ ms. The SNR gains obtained by the two methods are approximately the same.

TABLE X
SEPARATION OF CONVOLUTIVE MIXTURES CONSISTING OF FOUR SIGNALS.

# seg.	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Freq.
		$\tau_C = 200$		
0	–	–	–	0/50
1	–	–	–	0/50
2	–	–	–	0/50
3	56.30	45.74	6.22	29/50
4	65.21	49.85	5.57	21/50
		$\tau_C = 250$		
0	–	–	–	0/50
1	7.65	93.32	-5.20	1/50
2	45.61	49.19	6.73	1/50
3	56.42	48.90	6.01	30/50
4	62.90	50.32	5.62	18/50
		$\tau_C = 300$		
0	–	–	–	0/50
1	–	–	–	0/50
2	29.11	53.02	5.38	4/50
3	57.68	47.12	6.05	32/50
4	64.58	51.00	5.58	14/50
		$\tau_C = 350$		
0	–	–	–	0/50
1	–	–	–	0/50
2	36.86	53.85	5.56	9/50
3	54.83	47.63	5.97	30/50
4	65.02	49.55	5.71	11/50
		$\tau_C = 400$		
0	–	–	–	0/50
1	–	–	–	0/50
2	41.86	52.88	5.40	7/50
3	54.71	48.09	5.92	31/50
4	64.16	50.06	5.56	12/50

F. Comparison with other methods

Several other methods have been proposed for separation of an arbitrary number of speech mixtures with only two microphones by employing binary T-F masking [8], [24], [66]. In [24], speech signals were recorded binaurally and the interaural time difference (ITD) as well as the interaural intensity difference (IID) are extracted. The speech signals are separated by clustering in the joint ITD-IID domain. Separation results for three-source mixtures are given. An SNR gain of almost 14 dB is achieved. The gain also depends on the arrival directions of the source signals. Similarly, in the DUET algorithm described in [8], speech signals are separated by clustering speech signals in the amplitude/phase domain. In [8], the DUET algorithm was evaluated with synthetic anechoic mixtures, where amplitude and delay values are artificially chosen, as well as real reverberant recordings. These methods also have the advantage that the number of sources in the mixture need not be known in advance. In [24], the 128 frequency channels are (quasi) logarithmically distributed with center frequencies in the range of 80 Hz and 5000 Hz, while the frequency channels are linearly distributed in our proposed method and in [8] with a much higher frequency resolution.

In [40], the mask estimation is based on direction-of-arrival (DOA) techniques combined with ICA. The DOA technique is used to subtract $N - M$ sources, and the ICA algorithm is applied to the remaining M sources in the mixture. The method may be applied with binary masks, but in order to

reduce musical noise, more continuous masks based on the directivity patterns have been applied. The method is shown for separation of mixtures containing up to four speech signals. In contrast to [40], our method separates speech mixtures by iteratively extracting individual source signals. Similar to other multi-microphone methods our method relies on spatially different source locations, but unlike the previous methods, our method uses ICA to estimate the binary masks by iteratively estimating independent subsets of the mixtures. While methods based on DOA may sweep all possible directions in order to estimate the null directions, our proposed ICA technique automatically steers the nulls. Our approach can be used to iteratively steer the nulls in settings with more sources than microphones. In [41], binary masks are also found based on the ICA outputs. Our method differs from the method in [41] for our method is able to segregate more sources than mixtures.

Another method for extraction of multiple speakers with only two microphones is presented in [67]. This method is based on localization of the source signals followed by a cancellation part where for each time frame different nulls are steered for each frequency. Simulations under anechoic conditions show subtraction of speech signals in mixtures containing up to six equally loud source signals. In [67] the SNR is found with the original signals as ground truth. An SNR gain of 7–10 dB was reported. Our method gives a significantly higher ΔSNR.

The microphone placement is different in our method compared to the microphone placement in the DUET algorithm [8]. Therefore, in order to provide a fair comparison between our proposed and the DUET algorithm, we have implemented the DUET algorithm for demixing approximately W-disjoint orthogonal sources by following the stepwise description in [8].

1) Comparison with DUET in the instantaneous case: The DUET algorithm has been applied to the same set of instantaneous mixtures that were used in Table V and VI. The results of the DUET algorithm for separation of 3–6 sources are reported in Table XI. When comparing the separation results in Table XI with the results from our proposed method in Table V and VI, it can be seen that our proposed method gives a better ΔSNR. Note that our ΔSNR is different from the signal-to-interference ratio used in [8] and tends to be more stringent. Furthermore, our method is better at estimating the exact number of sources, as the Correct # column indicates. The histogram smoothing parameter in the DUET algorithm provides a delicate trade-off. If the histogram is smoothed too much, it results in sources that merge together. If the histogram is smoothed too little, erroneous peaks appear resulting in too high an estimate of the number of sources. The best performing setting of the smoothing parameter is used in our implementation.

2) Comparison with DUET for convolutive mixtures: The DUET algorithm has been applied to the same synthetic reverberant data set that was used in Section V-E. The separation performance can be found in Table XII. When comparing the results of the first parte in Table X and Table XII we find that the performance of the DUET algorithm and our proposed method is generally similar. Both algorithms have difficulties in finding the exact number of sources under reverberant conditions. The DUET is able to extract all four sources in 43 of the 50 experiments, while our method is able to extract all sources in 21 of the 50 experiments. The lower number of extracted sources in our proposed method is caused by our merging criterion which often tends to merge different sources. On the other hand, the SNR gain is a little higher for our method. In the remaining 29 experiments we are able to segregate three of the four sources, again with a higher SNR gain than the DUET algorithm.

TABLE XI
EVALUATION OF THE DUET ALGORITHM WITH RANDOM INSTANTANEOUS MIXTURES CONSISTING OF N SIGNALS.

N	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
3	26.61	20.04	-3.94	3.17	7.11	11/50
4	36.44	23.21	-5.77	2.04	7.63	20/50
5	39.42	22.95	-7.25	1.73	8.98	10/50
6	52.80	40.97	-8.20	0.30	8.51	1/50

TABLE XII
SEPARATION OF CONVOLUTIVE MIXTURES CONSISTING OF FOUR SIGNALS WITH THE DUET ALGORITHM.

# seg.	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Freq.
0	–	–	–	0/50
1	–	–	–	0/50
2	–	–	–	0/50
3	65.28	29.92	5.80	7/50
4	82.56	37.79	5.55	43/50

In summary, our comparison with DUET suggests that the proposed method produces better results for instantaneous mixtures and comparable results for convolutive mixtures. By listening to our results and those published in [8], the quality of our results seems at least as good as the quality of the separated signals of [8]. In terms of computational complexity, our method depends on the number of sources in the mixtures, whereas the complexity of the DUET algorithm mainly depends on the histogram resolution. We have chosen a histogram resolution of 101×101 and a smoothing kernel of size 20×20. With this histogram resolution, the DUET algorithm and our proposed method take comparable amounts of computing time, for convolutive mixtures about 20 minutes per mixture on average on an HP 320 server. For the instantaneous case, our algorithm is faster; for example, with three sources, it takes about 4:30 min ($\tau = 1$) and 3:40 min ($\tau = 2$) to segregate all the sounds from a mixture, and about 10 min ($\tau = 1$) and 7 min ($\tau = 2$) to segregate all the sounds when the instantaneous mixture consists of seven sources.

VI. DISCUSSION

In this paper directional microphones placed at the same location are assumed. This configuration allows the mixing matrix to be delay-less, and any standard ICA algorithm can therefore be applied to the problem. The configuration keeps the problem simple and still realistic. As shown in Section V-D, the algorithm may still be applied to delayed mixtures without significant changes. Alternatively, the ICA algorithm can be modified in order to separate delayed mixtures (see e.g. [4]). Since beamformer responses are used to de-

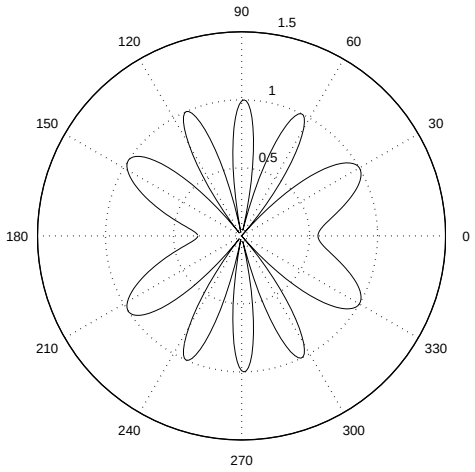


Fig. 12. A typical high-frequency microphone response. The response is given for the frequency of 4000 Hz, and a distance of 20 cm between the microphones. The half-wavelength at 4000 Hz is $\lambda/2 = 4.25$ cm. Since four whole half-wavelengths fit between the microphones, four nulls appear in the interval $0^\circ \leq \theta \leq 180^\circ$. Such a beampattern cannot efficiently be used to estimate the binary mask.

termine the binary masks, the microphone distance cannot be too big. If the distance between the microphones is greater than half the wavelength, spatial aliasing occurs, and frequency-dependent null directions and sidelobes occur. An example of such multiple null directions and sidelobes is shown in Fig. 12. Therefore, for large microphone distances, the performance is expected to decrease, especially at high frequencies. A solution to this problem could be to use the envelope of the mixed high-frequency signal as ICA input directly.

By only using instantaneous ICA in the reverberant case, we assume that the sources can be divided into many independent components that can be merged afterwards. However, this assumption has some limitations. Sometimes, the independent components are very sparse, and hence it is difficult to apply reliable grouping. A way to better cope with this problem and the delays may be to apply a convolutive separation algorithm instead of an instantaneous separation step. Still, we believe it is an advantage to use instantaneous source separation compared to convolutive source separation because it is computationally much simpler - it only has four values to estimate, whereas convolutive ICA has thousands of filter coefficients to estimate.

When binary time-frequency masks are used, artifacts (musical noise) are sometimes audible in the segregated signals, especially when the masks are sparse. The musical noise degrades the perceptual quality of the segregated signal. Musical noise is caused by several factors. The binary mask can be regarded as a time-variant gain function multiplied to the mixture in the frequency domain. This corresponds to a circular convolution in the time domain. Therefore artifacts due to aliasing occur. From an auditory point of view, musical noise appears when separated T-F regions are isolated from each other. As a result, the sound of such an isolated region becomes an audible tone, which does not group with the other sounds in the auditory scene. In order to reduce musical noise, it has been suggested to use continuous masks [40]. By listening to the signals, we have observed that a mask created by combining masks produced with different thresholds and weighted by the thresholds results in less musical artifacts. In our case, a more graded mask could be obtained by finding masks using different parameters τ and weighting the T-F units of the masks with the corresponding thresholds or simply by smoothing the binary mask in time and in frequency.

Our method has also been applied to separate stereo music. Stereo signals are often constructed by applying different gains to the different instruments on the two channels. Sometimes stereo signals are created with directional microphones placed at the same location with an 90° angle between the directional patterns. Our method is able to segregate single instruments or vocal sounds from the stereo music mixture [44].

In the evaluation the source directions are limited to seven different directions uniformly distributed on a half-circle. In a real environment, speech signals may arrive from closer directions. Also, with only two microphones, it is not possible to distinguish the two half-planes divided by the microphone array. If two arrival angles become too close, the source signals can no longer be segregated and two spatially close sources may be considered as a single source by the stopping criterion. When two sources are treated as a single source depends on the number of sources in the mixture. In the evaluation, it becomes harder to segregate all N sources as N increases. Also the level of background/microphone noise influences the spatial resolution.

Several issues in our proposed method need further investigation. Different criteria have been proposed in order to decide when the iterations should stop and when different binary masks should be merged. These criteria need to set many parameters and many experiments are needed on order to optimize these parameters. Furthermore, the optimal parameters most likely depend on a given setting, e.g. the number of sources in the mixture or the amount of reverberation. The stopping criterion was proposed for the instantaneous mixing case but applied to reverberant mixtures too. A more robust stopping criterion in the convolutive case would be a subject for future work. Our grouping criterion in the convolutive case is based on correlation between different envelopes. One could interpret the grouping problem as a problem similar to a frequency permutation problem known in blind source separation (see e.g. [68]). The merging criterion may be more reliable if it is combined with other cues, such as DOA information.

VII. Conclusion

We have proposed a novel method for separating instantaneous and anechoic mixtures with an arbitrary number of speech signals of equal power with only two microphones. We have dealt with underdetermined mixtures by applying ICA to produce independent subsets. The subsets are used to estimate binary T-F masks, which are then applied to separate original mixtures. This iterative procedure continues until the independent subsets consist of only a single source. The segregated signals are further improved by merging masks from

correlated subsets. Extensive evaluation shows that mixtures of up to seven speech signals under anechoic conditions can be separated. The estimated binary masks are close to the ideal binary masks. The proposed framework has also been applied to speech mixtures recorded in a reverberant room. We find that instantaneous ICA applied iteratively in the time domain can be used to segregate convolutive mixtures. The performance of our method compares favorably with other methods for separation of underdetermined mixtures. Because the sources are iteratively extracted from the mixture the number of sources does not need to be assumed in advance; except for reverberant mixtures our method gives a good estimate of the number of sources. Further, stereo signals are maintained throughout the processing.

APPENDIX
DIRECTIONAL GAINS

The two directional gain patterns can be approximated from two closely-spaced omnidirectional microphones. The directional response from two microphones can be written as

$$r(\theta) = s_1 e^{j\frac{kd}{2}\cos(\theta)} + s_2 e^{-j\frac{kd}{2}\cos(\theta)}, \tag{25}$$

where s_1 and s_2 are the microphone sensitivities. $k = 2\pi/\lambda = 2\pi f/c$ is the wave number. f is the acoustic frequency and $c = 343$ m/s is the speed of sound traveling in the air at 20°C. θ is the angle between the microphone array line and the source direction of arrival and d is the distance between the two microphones. If $kd \ll 1$, the microphone response can be approximated by [69]

$$r(\theta) \approx A + B\cos(\theta), \tag{26}$$

where $A = s_1 + s_2$ and $B = \frac{j}{kd}(s_1 - s_2)$. Here,

$$s_1 = \frac{1}{2}A - \frac{j}{kd}B \tag{27}$$

$$s_2 = \frac{1}{2}A + \frac{j}{kd}B. \tag{28}$$

In the Laplacian domain, $s = j\omega$, we have

$$s_1 = \frac{1}{2}A + \frac{c}{sd}B \tag{29}$$

$$s_2 = \frac{1}{2}A - \frac{c}{sd}B. \tag{30}$$

For discrete signals, we use the bilinear transform [70]

$$s = 2f_s\frac{1 - z^{-1}}{1 + z^{-1}}, \tag{31}$$

where f_s is the sampling frequency. The two discrete microphone sensitivities are therefore

$$s_1 = \frac{(Af_sd + cB) + (cB - Af_sd)z^{-1}}{2f_sd(1 - z^{-1})} \tag{32}$$

$$s_2 = \frac{(Af_sd - cB) - (cB + Af_sd)z^{-1}}{2f_sd(1 - z^{-1})} \tag{33}$$

It can be seen that the denominators in (32) and (33) have a root on the unit circle. In order to ensure stability, we modify the denominator with a factor λ so that

$$s_1 = \frac{(Af_sd + cB) + (cB - Af_sd)z^{-1}}{2f_sd(1 - \lambda z^{-1})} \tag{34}$$

$$s_2 = \frac{(Af_sd - cB) - (cB + Af_sd)z^{-1}}{2f_sd(1 - \lambda z^{-1})} \tag{35}$$

We choose $\lambda = 0.75$. λ controls the gain that amplifies the low frequencies. The choice of λ is not very important, because the signals are used for comparison only.

In order to obtain the directional patterns in Fig. 1 we can find A and B by solving (26) for two different gains. For $r(0) = 1$ and $r(\pi) = 0.5$, we obtain $A = 0.75$ and $B = 0.25$. For $r(0) = 0.5$ and $r(\pi) = 1$, we obtain $A = 0.75$ and $B = -0.25$.

ACKNOWLEDGMENT

The work was performed while M.S.P. was a visiting scholar at The Ohio State University Department of Computer Science and Engineering. M.S.P was supported by the Oticon Foundation. M.S.P. and J.L. are partly also supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778. D.L.W. was supported in part by an AFOSR grant (FA9550-04-1-0117) and an AFRL grant (FA8750-04-0093).

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, September 1953.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, September 2005.

[3] L. K. Hansen, "Blind separation of noisy image mixtures," in *Advances in Independent Component Analysis, Perspectives in Neural Computing*, M. Girolami, Ed. Springer-Verlag, 2000, ch. 9, pp. 165–187.

[4] K. Torkkola, "Blind separation of delayed and convolved sources," in *Unsupervised Adaptive Filtering, Blind Source Separation*, S. Haykin, Ed. Wiley, John and Sons, Incorporated, January 2000, vol. 1, ch. 8, pp. 321–375.

[5] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '03)*, vol. 6, 2003, pp. VI–581–584.

[6] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 888–893, July 2002.

[7] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, September 2003, pp. 271–274.

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

[9] R. K. Olsson and L. K. Hansen, "Blind separation of more sources than sensors in convolutive mixtures," in *ICASSP*, 2006. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4321

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[11] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing, Elsevier*, vol. 24, no. 1, pp. 1–10, 1991.

[12] P. Comon, E. Moreau, and L. Rota, "Blind separation of convolutive mixtures: A contrast-based joint diagonalization approach," in *3rd Int. Conf. Independent Component Analysis*, San Diego, CA, December 9–13 2001.

[13] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[14] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. (4–5), pp. 411–430, 2000.

[15] L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[16] T. Kolenda, L. Hansen, and J. Larsen, "Signal detection using ica: Application to chat room topic spotting," *In proc. ICA'2001*, pp. 540–545, 2001. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?826

[17] L. K. Hansen, J. Larsen, and T. Kolenda, "On independent component analysis for multimedia signals," in *Multimedia Image and Video Processing*, L. Guan, S. Kung, and J. Larsen, Eds. CRC Press, 2000, pp. 175–199. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?627

[18] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[19] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[20] D. Luengo, I. Santamaria, L. Vielva, and C. Pantaleon, "Underdetermined blind separation of sparse sources with instantaneous and convolutive mixtures," in *IEEE XIII Workshop on Neural Networks for Signal Processing*, 2003, pp. 279–288.

[21] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *ICA 2006*, 2006, pp. 536–543.

[22] D. B. Rowe, "A Bayesian approach to blind source separation," *Journal of Interdisciplinary Mathematics*, vol. 5, no. 1, pp. 49–76, 2002.

[23] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, pp. 2353–2362, 2001.

[24] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, October 2003.

[25] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete bss for convolutive mixtures based on hierarchical clustering," in *Proc. ICA'2004*, Granada, Spain, September 22–24 2004, pp. 652–660.

[26] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," Tech. Rep., April 4 2005, to appear in IJIST (International Journal of Imaging Systems and Technology), special issue on Blind Source Separation and Deconvolution in Imaging and Image Processing.

[27] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of Eurospeech03*, Geneva, September 2003, pp. 1009–1012.

[28] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP'2000*, vol. V, Istanbul, Turkey, June 2000, pp. 2985–2988.

[29] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley & IEEE Press, 2006.

[30] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[31] A. S. Bregman, *Auditory Scene Analysis*, 2nd ed. MIT Press, 1990.

[32] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[33] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. Academic Press, 2003.

[34] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, September 2004.

[35] S. Roweis, "One microphone source separation," in *NIPS'00*, 2000, pp. 793–799.

[36] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2014–2023, Nov. 2006.

[37] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing, Elsevier*, vol. 87, no. 8, pp. 1833–1847, 2007.

[38] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 34, no. 4, pp. 1763–1773, August 2004.

[39] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP2005*, vol. III, March 18–23 2005, pp. 81–84.

[40] ——, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA," in *Proc. ICA'2004*, September 22–24 2004, pp. 898–905.

[41] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. ICA'2004*, Granada, Spain, September 22–24 2004, pp. 832–839.

[42] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1592–1604, July 2007.

[43] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Overcomplete blind source separation by combining ICA and binary time-frequency masking," in *Proceedings of the MLSP workshop*, Mystic, CT, USA, September 28–30 2005. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3894

[44] M. S. Pedersen, T. Lehn-Schiøler, and J. Larsen, "BLUES from music: BLind Underdetermined Extraction of Sources from Music," in *ICA2006*, 2006, pp. 392–399. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4060

[45] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *ICA 2006*, 2006, pp. 674–681. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4068

[46] M. Ito, Y. Takeuchi, T. Matsumoto, H. Kudo, M. Kawamoto, T. Mukai, and N. Ohnishi, "Moving-source separation using directional microphones," in *Proc. ISSPIT'2002*, December 2002, pp. 523–526.

[47] G. Cauwenberghs, M. Stanacevic, and G. Zweig, "Blind broadband source localization and separation in miniature sensor arrays," in *IEEE Int. Symp. Circuits and Systems (ISCAS'2001)*, vol. 3, May 6–9 2001, pp. 193–196.

[48] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays*, ser. Digital Signal Processing. Springer, 2001.

[49] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamformers," in *Consistent & Reliable Acoustic Cues for Sound Analysis (CRAC)*, September 2001.

[50] R. D. Patterson, "The sound of a sinusoid: Spectral models," *J. Acoust. Soc. Am.*, vol. 96, pp. 1409–1418, May 1994.

[51] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *J. Audio Eng. Soc*, vol. 48, no. 11, pp. 1011–1031, November 2000.

[52] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[53] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind sourc separation," in *Proc. 2nd IEEE Int. Workshop Indep. Compon. Anal. Signal Separation*, Helsinki, Finland, June 2000, pp. 215–220.

[54] W. M. Hartmann, *Signals, Sound, and Sensation*. Springer, 1998.

[55] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. Baltimore: The Johns Hopkins University Press, 1996.

[56] D. P. W. Ellis, "Evaluating speech separation systems," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, ch. 20, pp. 295–304.

[57] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures [demo]," January 2006. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4400

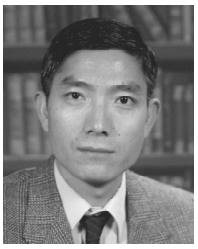
[58] ——, "Two-microphone separation of speech mixtures [source code]," January 2006. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4399

[59] H. Nielsen, "UCMINF - an algorithm for unconstrained, nonlinear optimization," IMM, Technical University of Denmark, Tech. Rep. IMM-TEC-0019, 2001. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?642

[60] T. Kolenda, S. Sigurdsson, O. Winther, L. K. Hansen, and J. Larsen, "DTU:toolbox," Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark, 2002. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4043

[61] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, January 1978.
[62] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, December 1993.
[63] "ICA central." [Online]. Available: http://www.tsi.enst.fr/icacentral/algos.html
[64] "Brüel & Kjær Head and Torso Simulator, Type 4128."
[65] F. Asano, Y. Motomura, H. Asoh, and T. Matsui, "Effect of PCA in blind source separation," in *Proceedings of the Second International Workshop on ICA and BSS*, P. Pajunen and J. Karhunen, Eds., Helsinki, Finland, June 19–22 2000, pp. 57–62.
[66] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4040–4051, 2006.
[67] C. Liu, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust Soc. Amer.*, vol. 110, pp. 3218–3231, 2001.
[68] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, September 2004.
[69] S. C. Thompson, "Directional patterns obtained from two or three microphones," Knowles Electronics," Technical Report, September 29 2000.
[70] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*. Prentice-Hall, 1996.

Michael Syskind Pedersen received the M.Sc. degree in 2003 from the Technical University of Denmark (DTU). In 2006 he obtained his Ph.D. degree from the department of Informatics and Mathematical Modelling (IMM) at DTU. In 2005 he was a Visiting Scholar at the Department of Computer Science & Engineering at The Ohio State University, Columbus, OH. Michael's main areas of research are multi-microphone audio processing and noise reduction. Since 2001 Michael has been employed with the hearing aid company Oticon.



DeLiang Wang (M'90-SM'01-F'04) received the B.S. degree in 1983 and the M.S. degree in 1986 from Peking (Beijing) University, Beijing, China, and the Ph.D. degree in 1991 from the University of Southern California, Los Angeles, CA, all in computer science.

From July 1986 to December 1987 he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science & Engineering and the Center for Cognitive Science at The Ohio State University, Columbus, OH, where he is currently a Professor. From October 1998 to September 1999, he was a visiting scholar in the Department of Psychology at Harvard University, Cambridge, MA.

Dr. Wang's research interests include machine perception and neurodynamics. He is a recipient of the 1996 U.S. Office of Naval Research Young Investigator Award. Dr. Wang served as the President of the International Neural Network Society in 2006.

Jan Larsen received the M.Sc. and Ph.D. degrees in electrical engineering from the Technical University of Denmark (DTU) in 1989 and 1994. Dr. Larsen is currently Associate Professor of Digital Signal Processing at Informatics and Mathematical Modelling, DTU. Jan Larsen has authored and co-authored more than 100 papers and book chapters within the areas of nonlinear statistical signal processing, machine learning, neural networks and datamining with applications to biomedicine, monitoring systems, multimedia, and webmining.

He has participated in several national and international research programs, and has served as reviewer for many international journals, conferences, publishing companies and research funding organizations. Further he took part in conference organizations, among others, the IEEE Workshop on Machine Learning for Signal Processing (formerly Neural Networks for Signal Processing) 1999–2007. He is past chair of the IEEE Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2005-2007), and chair of IEEE Denmark Section's Signal Processing Chapter (2002-). He is a senior member of The Institute of Electrical and Electronics Engineers. Other professional committee participation includes: Member of the Technical Committee 14: Signal Analysis for Machine Intelligence of the International Association for Pattern Recognition, 2006-; Steering committee member of the Audio Signal Processing Network in Denmark, 2006-. Editorial Board Member of Signal Processing, Elsevier, 2006-2007; and guest editorships involves IEEE Transactions on Neural Networks; Journal of VLSI Signal Processing Systems; and Neurocomputing.

Ulrik Kjems was born March 1971, he obtained his Master of Science degree in Electrical Engineering at the Technical University of Denmark (DTU) February 1995, and his Ph.D. at the Section for Digital Signal Processing at the Department of Mathematical Modelling, also at DTU. His Ph.D. research areas have been functional and anatomical brain scans and deformable models of the brain, followed by 2 year post doc position working with statistical models of functional activation patterns in brain scans. From 2000 he has been a design engineer at Oticon Denmark, developing audio processing algorithms, working with source separation, beam forming, noise reduction and compression algorithms for hearing aids.