Two-Microphone Separation of Speech Mixtures

Pedersen, Michael Syskind; Wang, DeLiang; Larsen, Jan; Kjems, Ulrik

Published in:
IEEE Transactions on Neural Networks

Link to article, DOI:
10.1109/TNN.2007.911740

Publication date:
2008

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
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 $N_i = f_sT$ 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.,

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

where $\mathbf{A}$ is an $M \times N$ mixing matrix. $\mathbf{\nu}(n)$ 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 $\hat{\mathbf{y}}(n)$ of the original source signals can be found by applying an (pseudo) inverse linear operation, i.e.,

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

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

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

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 under-determined (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

$$S_i(\omega, t)S_j(\omega, t) = 0, \forall i \neq j, \forall \omega, t,$$

where $t$ is the time frame index and $\omega$ 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 $N - M$ sources by masking and then applying ICA in order to separate the remaining $M$ sources. In [41], ICA has been used the other way around. Here, ICA is applied to separate two sources 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 $(2 \times 2)$ 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 Appendix A, 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 $\theta$ 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
corresponds to a given direction, a given set of microphone
 gains corresponds to a certain direction, and the mixing system
 is given by

\[
A(\theta) = \begin{bmatrix}
  r_1(\theta_1) & \cdots & r_1(\theta_N) \\
  r_2(\theta_1) & \cdots & r_2(\theta_N)
\end{bmatrix}.
\]  \hspace{1cm} (5)

For the instantaneous case, the separation matrix \( W \) 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 \( y(n) = [y_1(n), y_2(n)]^T \). Since the number of
separated signals obtained by (2) is smaller than the number
of source signals, \( y \) does not contain the separated signals.
Instead, if the noise term is disregarded, \( y \) is another linear
superposition of the source signals, i.e.

\[
y = Gs,
\]  \hspace{1cm} (6)

where the weights are given by \( G = WA \) instead of just \( A \)
as in (1). Thus, \( G \) just corresponds to another weighting of
each of the source signals depending on \( \theta \). 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 \( G(\theta) \) in the underdetermined case.
The response of the \( m \)'th output is given by \( g_{mn}(\theta) = [w_m^T, a(\theta)] \),
where \( w_m \) is the separation vector from the \( m \)'th output and
\( a(\theta) = [r_1(\theta), r_2(\theta)]^T \) is the mixing vector for the arrival
direction \( \theta \) [48]. By varying \( \theta \) 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,
\( Y(\omega) = G(\omega)S(\omega) \).

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. 1. The two directional microphone responses are shown as function of
the direction \( \theta \).

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 \( 2 \times 2 \) 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.
The two estimated masks are then applied in the T-F domain to the original input binary mask to prevent re-introduction of already masked T-F units. The four masked signals and corresponding T-F units of the two T-F signals and multiplying by the method to solve the scaling ambiguity. As proposed in [10], ICA outputs, it is necessary to solve the scaling problem. In the algorithm are arbitrarily scaled (step C) 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 \( y_2 \). \( y_1 \) and \( y_2 \) 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.

\[
BM_1(\omega, t) = \begin{cases} 
1, & \text{if } |Y_1(\omega, t)| \geq \tau |Y_2(\omega, t)|; \\
0, & \text{otherwise.}
\end{cases} \quad (9)
\]

\[
BM_2(\omega, t) = \begin{cases} 
1, & \text{if } |Y_2(\omega, t)| \geq \tau |Y_1(\omega, t)|; \\
0, & \text{otherwise.}
\end{cases} \quad (10)
\]

where \( \omega \) 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):

\[
BM_1(\omega, t) = \begin{cases} 
1, & \text{if } |Y_1(\omega, t)| > \tau |Y_2(\omega, t)|; \\
0, & \text{otherwise.}
\end{cases} \quad (9)
\]

\[
BM_2(\omega, t) = \begin{cases} 
1, & \text{if } |Y_2(\omega, t)| > \tau |Y_1(\omega, t)|; \\
0, & \text{otherwise.}
\end{cases} \quad (10)
\]

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 \( \hat{p} \).
Reestimate signals from $KF$ decomposition, where the frequency spacing is logarithmic scale. The frequency resolution at the low frequencies is much representation in the human ear is closer to a logarithmic accord well with human perception of sounds. The frequency a linear frequency scale. A linear frequency scale does not are reconstructed in the time domain by the inverse STFT (step mask. After the masks have been applied to the signals, they are segregated in the time domain and other speakers are attenuated by the other processing, some of the speech signals are poor quality are merged with these masks. A masks output signals which are correlated with signals of improve the separated signals. Masks of output signals which are correlated signal. The outputs from the main algorithm are further processed in order to segregated signals, and poor quality signals as $\hat{\rho}$. 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 $\hat{\delta}$. 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 sources. 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 segregated 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 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, $BM^{(0)}$ 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 $\hat{\rho}$. 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 $\hat{\delta}$. 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 segregated 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 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, $BM^{(0)}$ 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 $\hat{\rho}$. 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 $\hat{\delta}$. 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 segregated 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 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].
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 $\tau$ 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.

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 $\tau_{C2}$, 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.
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 \( \mathbf{x} = \mathbf{A} \mathbf{s} \) with \( \mathbf{s} = [\alpha_3 s_3(n-k_3), s_1(n), \alpha_2 s_2(n-k_2), s_3(n), \alpha_1 s_1(n-k_1)]^T \) and

\[
\mathbf{A}(\theta) = \begin{bmatrix}
\tau_1(\theta_1) & \cdots & \tau_1(\theta_6) \\
\tau_2(\theta_1) & \cdots & \tau_2(\theta_6)
\end{bmatrix}.
\]

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 $E$ of the signal $x(n)$ can be calculated as

$$E(x(n)) = |x(n) + jH(x(n))|,$$  

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

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

where

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

The above values of $\alpha$ 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 $\hat{E}(x(0)) = 0$.

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.

### 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

$$R_{xx} = \langle xx^T \rangle = \frac{1}{N_s} xx^T,$$

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

$$R_{xx} = \langle (As + \nu)(As + \nu)^T \rangle = A\langle ss^T \rangle A^T + \langle \nu \nu^T \rangle = \Psi + \sigma^2 I,$$

where $\Psi = A\langle ss^T \rangle A^T$ 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 $\tau_c$, i.e.

$$\text{cond}(R_{xx}) > \tau_c.$$  

A high condition number indicates that the matrix is close to being singular. Since $R_{xx}$ is symmetric and positive definite, $\text{cond}(R_{xx}) = \max \text{eig}(R_{xx}) / \min \text{eig}(R_{xx})$, where $\text{eig}(R_{xx})$ is the vector of eigenvalues of $R_{xx}$. 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 $\tau_E$, which is measured in $\text{dB}$.

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:

\[
P_{EL} = \frac{\sum e_1^2(n)}{\sum I^2(n)} \quad (21)
\]

\[
P_{NR} = 1 - \frac{\sum e_2^2(n)}{\sum O^2(n)} \quad (22)
\]

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

\[
\text{SNR}_o = 10 \log_{10} \left[ \frac{\sum I^2(n)}{\sum (I(n) - O(n))^2} \right] \quad (23)
\]

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 disproportionally 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

\[
\Delta \text{SNR} = \text{SNR}_o - \text{SNR}_i \quad (24)
\]

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 \( f_s = 10 \text{ kHz} \). 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 \( 30^\circ \). 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**

<table>
<thead>
<tr>
<th>$\tau_E$</th>
<th>$N = 4$</th>
<th>$\tau_C$</th>
<th>$N = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2000</td>
<td>3000</td>
<td>4000</td>
</tr>
<tr>
<td>15</td>
<td>15.45</td>
<td>15.34</td>
<td>15.24</td>
</tr>
<tr>
<td>10/10</td>
<td>6/10</td>
<td>5/10</td>
<td>5/10</td>
</tr>
<tr>
<td>20</td>
<td>15.34</td>
<td>15.23</td>
<td>15.18</td>
</tr>
<tr>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
<td>8/10</td>
</tr>
<tr>
<td>4/10</td>
<td>4/10</td>
<td>5/10</td>
<td>5/10</td>
</tr>
</tbody>
</table>

$\Delta \text{SNR}$ 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 $\tau$ 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, $\tau_C1$ and $\tau_C2$ 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 $\tau_C1$ is increased, some signals may not be merged even though they mainly contain the same source. If $\tau_C2$ 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 $\tau_C2$ is increased, the masks become sparser, and more artifacts occur. If $\tau_C2$ 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 $\tau_C$ and $\tau_E$ has been evaluated. $\tau_C$ has been evaluated for the values 2000, 3000 and 4000. Likewise, $\tau_E$ has been evaluated for the values 15, 20 and 25 dB. For each pair of $\tau_C$ and $\tau_E$ 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, $\Delta \text{SNR}$ 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 $\Delta \text{SNR}$ 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.

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.
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 cases of two sources. 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. 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 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 V

**Evaluation with random instantaneous mixtures consisting of $N$ signals.**

<table>
<thead>
<tr>
<th>$N$</th>
<th>$P_{EL}$ (%)</th>
<th>$P_{NR}$ (%)</th>
<th>SNR$_i$</th>
<th>SNR$_o$</th>
<th>$\Delta$SNR</th>
<th>Correct #</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.01</td>
<td>2.00</td>
<td>0</td>
<td>18.92</td>
<td>18.92</td>
<td>47/50</td>
</tr>
<tr>
<td>3</td>
<td>2.99</td>
<td>4.86</td>
<td>-3.95</td>
<td>12.50</td>
<td>16.45</td>
<td>46/50</td>
</tr>
<tr>
<td>4</td>
<td>6.07</td>
<td>8.62</td>
<td>-5.98</td>
<td>9.26</td>
<td>15.23</td>
<td>46/50</td>
</tr>
<tr>
<td>5</td>
<td>10.73</td>
<td>13.02</td>
<td>-7.40</td>
<td>5.56</td>
<td>14.27</td>
<td>44/50</td>
</tr>
<tr>
<td>6</td>
<td>14.31</td>
<td>13.63</td>
<td>-8.39</td>
<td>5.25</td>
<td>13.64</td>
<td>44/50</td>
</tr>
<tr>
<td>7</td>
<td>18.34</td>
<td>22.43</td>
<td>-9.24</td>
<td>4.24</td>
<td>13.48</td>
<td>41/50</td>
</tr>
</tbody>
</table>

The parameter $\tau = 1$.

### Table VI

**Evaluation with random instantaneous mixtures consisting of $N$ signals.**

<table>
<thead>
<tr>
<th>$N$</th>
<th>$P_{EL}$ (%)</th>
<th>$P_{NR}$ (%)</th>
<th>SNR$_i$</th>
<th>SNR$_o$</th>
<th>$\Delta$SNR</th>
<th>Correct #</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.43</td>
<td>0.50</td>
<td>0</td>
<td>18.22</td>
<td>18.22</td>
<td>50/50</td>
</tr>
<tr>
<td>3</td>
<td>7.36</td>
<td>2.60</td>
<td>-3.96</td>
<td>11.10</td>
<td>15.06</td>
<td>46/50</td>
</tr>
<tr>
<td>4</td>
<td>12.26</td>
<td>4.17</td>
<td>-5.89</td>
<td>8.81</td>
<td>14.70</td>
<td>42/50</td>
</tr>
<tr>
<td>5</td>
<td>19.81</td>
<td>6.21</td>
<td>-7.32</td>
<td>6.59</td>
<td>13.91</td>
<td>40/50</td>
</tr>
<tr>
<td>6</td>
<td>25.91</td>
<td>8.81</td>
<td>-8.36</td>
<td>5.31</td>
<td>13.67</td>
<td>23/50</td>
</tr>
<tr>
<td>7</td>
<td>30.52</td>
<td>11.86</td>
<td>-9.12</td>
<td>3.00</td>
<td>13.46</td>
<td>4/50</td>
</tr>
</tbody>
</table>

The parameter $\tau = 2$.

### Table VII

**Evaluation of separation performance as function of the signal length $T$.**

<table>
<thead>
<tr>
<th>$T$</th>
<th>$P_{EL}$ (%)</th>
<th>$P_{NR}$ (%)</th>
<th>SNR$_i$</th>
<th>SNR$_o$</th>
<th>$\Delta$SNR</th>
<th>Correct #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.53</td>
<td>5.83</td>
<td>-6.38</td>
<td>9.44</td>
<td>15.83</td>
<td>34/50</td>
</tr>
<tr>
<td>2</td>
<td>7.85</td>
<td>8.23</td>
<td>-5.98</td>
<td>9.00</td>
<td>14.88</td>
<td>43/50</td>
</tr>
<tr>
<td>3</td>
<td>6.87</td>
<td>9.69</td>
<td>-6.04</td>
<td>8.80</td>
<td>14.85</td>
<td>46/50</td>
</tr>
<tr>
<td>4</td>
<td>7.57</td>
<td>9.05</td>
<td>-6.04</td>
<td>8.81</td>
<td>14.86</td>
<td>46/50</td>
</tr>
<tr>
<td>5</td>
<td>6.07</td>
<td>8.62</td>
<td>-5.98</td>
<td>9.26</td>
<td>15.23</td>
<td>46/50</td>
</tr>
</tbody>
</table>

Instantaneous 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 $\Delta SNR$ is 16.57 dB, $P_{25} = 6.65\%$ and $P_{40} = 14.64\%$. When we compare to the more difficult case in Table V where all four speakers have equal loudness, we see that the average $\Delta SNR$ 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.

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 $5.2 \times 7.9 \times 3.5$ m and the distance between the microphones and the loudspeakers were 2 m. Impulse responses from loudspeaker positions of $0^\circ$, $90^\circ$, $135^\circ$, and $180^\circ$ 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 $T_C = 2000$. Therefore we need much lower thresholds. The separation performance is found for different values of $T_C$. 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

---

**TABLE VIII**

<table>
<thead>
<tr>
<th>Noise (dB)</th>
<th>$P_{25}$ (%)</th>
<th>$P_{40}$ (%)</th>
<th>$SNR$</th>
<th>$SNR_0$</th>
<th>$\Delta SNR$</th>
<th>Correct #</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
<td>15.29</td>
<td>15.52</td>
<td>-6.51</td>
<td>5.95</td>
<td>12.43</td>
<td>19/50</td>
</tr>
<tr>
<td>-20</td>
<td>7.42</td>
<td>10.26</td>
<td>-6.02</td>
<td>8.37</td>
<td>14.39</td>
<td>45/50</td>
</tr>
<tr>
<td>-30</td>
<td>6.24</td>
<td>8.53</td>
<td>-5.99</td>
<td>9.27</td>
<td>15.26</td>
<td>46/50</td>
</tr>
<tr>
<td>-40</td>
<td>6.23</td>
<td>8.72</td>
<td>-5.97</td>
<td>9.19</td>
<td>15.16</td>
<td>47/50</td>
</tr>
<tr>
<td>-50</td>
<td>6.39</td>
<td>8.15</td>
<td>-5.98</td>
<td>9.29</td>
<td>15.27</td>
<td>45/50</td>
</tr>
<tr>
<td>-60</td>
<td>6.04</td>
<td>8.62</td>
<td>-5.98</td>
<td>9.27</td>
<td>15.25</td>
<td>46/50</td>
</tr>
</tbody>
</table>

---

**TABLE IX**

<table>
<thead>
<tr>
<th>Mic. dist.</th>
<th>$P_{25}$ (%)</th>
<th>$P_{40}$ (%)</th>
<th>$\Delta SNR$</th>
<th>Correct #</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 1$ cm</td>
<td>7.63</td>
<td>8.84</td>
<td>14.83</td>
<td>17/50</td>
</tr>
<tr>
<td>Ideal</td>
<td>6.07</td>
<td>8.62</td>
<td>15.23</td>
<td>46/50</td>
</tr>
</tbody>
</table>

Anechoic mixtures consisting of four sources have been used.

---
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.

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

\begin{table}[h]
\centering
\caption{Separation of convolutive mixtures consisting of four signals.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\# seg. & $P_{\text{EL}}(\%)$ & $P_{\text{NR}}(\%)$ & $\Delta\text{SNR}$ & Freq. \\
\hline
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 \\
\hline
\hline
\hline
\hline
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Separation of convolutive mixtures consisting of four signals.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\# seg. & $P_{\text{EL}}(\%)$ & $P_{\text{NR}}(\%)$ & $\Delta\text{SNR}$ & Freq. \\
\hline
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 \\
\hline
\hline
\hline
\hline
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Separation of convolutive mixtures consisting of four signals.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\# seg. & $P_{\text{EL}}(\%)$ & $P_{\text{NR}}(\%)$ & $\Delta\text{SNR}$ & Freq. \\
\hline
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 \\
\hline
\hline
\hline
\hline
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Separation of convolutive mixtures consisting of four signals.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\# seg. & $P_{\text{EL}}(\%)$ & $P_{\text{NR}}(\%)$ & $\Delta\text{SNR}$ & Freq. \\
\hline
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 \\
\hline
\hline
\hline
\hline
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Separation of convolutive mixtures consisting of four signals.}
\begin{tabular}{|c|c|c|c|c|}
\hline
\# seg. & $P_{\text{EL}}(\%)$ & $P_{\text{NR}}(\%)$ & $\Delta\text{SNR}$ & Freq. \\
\hline
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 \\
\hline
\hline
\hline
\hline
\hline
\end{tabular}
\end{table}
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 ∆SNR.

The microphone placement is different in our method com-
pared 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 ∆SNR. Note that our ∆SNR 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 rever-
berant 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 part 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.

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 algo-

TABLE XI

<table>
<thead>
<tr>
<th>N</th>
<th>$P_{tr}(%)$</th>
<th>$P_{nr}(%)$</th>
<th>SNR$^i$</th>
<th>SNR$^o$</th>
<th>∆SNR</th>
<th>Correct #</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>26.61</td>
<td>20.04</td>
<td>-3.94</td>
<td>3.17</td>
<td>7.11</td>
<td>11/50</td>
</tr>
<tr>
<td>4</td>
<td>36.44</td>
<td>23.21</td>
<td>-5.77</td>
<td>2.04</td>
<td>7.63</td>
<td>20/50</td>
</tr>
<tr>
<td>5</td>
<td>39.42</td>
<td>22.95</td>
<td>-7.25</td>
<td>1.73</td>
<td>8.98</td>
<td>10/50</td>
</tr>
<tr>
<td>6</td>
<td>52.80</td>
<td>40.97</td>
<td>-8.20</td>
<td>0.30</td>
<td>8.51</td>
<td>1/50</td>
</tr>
</tbody>
</table>

TABLE XII

<table>
<thead>
<tr>
<th># seg.</th>
<th>$P_{tr}(%)$</th>
<th>$P_{nr}(%)$</th>
<th>∆SNR</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0/50</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0/50</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0/50</td>
</tr>
<tr>
<td>3</td>
<td>65.28</td>
<td>29.92</td>
<td>5.80</td>
<td>7/50</td>
</tr>
<tr>
<td>4</td>
<td>82.56</td>
<td>37.79</td>
<td>5.55</td>
<td>43/50</td>
</tr>
</tbody>
</table>

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 mix-
tures (see e.g. [4]). Since beamformer responses are used to de-
We believe it is an advantage to use instantaneous 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 signal, 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 $\tau$ 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\(^\circ\) 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

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

(25)

where $s_1$ and $s_2$ are the microphone sensitivities, $k = 2\pi/\lambda$ 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. $\theta$ 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]

$$ r(\theta) \approx A + B \cos(\theta), $$

(26)

where $A = s_1 + s_2$ and $B = \frac{f_s}{2}(s_1 - s_2)$. Here,

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

(27)

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

(28)

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

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

(29)

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

(30)

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

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

(31)

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

$$ s_1 = \frac{(Af_f d + cB) + (cB - Af_f d)z^{-1}}{2f_s d(1 - z^{-1})}, $$

(32)

$$ s_2 = \frac{(Af_f d - cB) - (cB + Af_f d)z^{-1}}{2f_s d(1 - z^{-1})}. $$

(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 $\lambda$ so that

$$ s_1 = \frac{(Af_f d + cB) + (cB - Af_f d)z^{-1}}{2f_s d(1 - \lambda z^{-1})}, $$

(34)

$$ s_2 = \frac{(Af_f d - cB) - (cB + Af_f d)z^{-1}}{2f_s d(1 - \lambda z^{-1})}. $$

(35)

We choose $\lambda = 0.75$. $\lambda$ controls the gain that amplifies the low frequencies. The choice of $\lambda$ 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**


[64] “Bruel & Kjaer Head and Torso Simulator, Type 4128.”


