OVERCOMPLETE BLIND SOURCE SEPARATION BY COMBINING ICA AND BINARY TIME-FREQUENCY MASKING

Michael Syskind Pedersen[1,2], *DeLiang Wang*[3], *Jan Larsen*[1] *and Ulrik Kjems*[2]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
[2]Oticon A/S, Strandvejen 58, DK-2900 Hellerup, Denmark
[3] Department of Computer Science and Engineering & Center for Cognitive Science,
The Ohio State University, Columbus, OH 43210-1277, USA

ABSTRACT

A limitation in many source separation tasks is that the number of source signals has to be known in advance. Further, in order to achieve good performance, the number of sources cannot exceed the number of sensors. In many real-world applications these limitations are too strict. We propose a novel method for overcomplete blind source separation. Two powerful source separation techniques have been combined, *independent component analysis* and *binary time-frequency masking*. Hereby, it is possible to iteratively extract each speech signal from the mixture. By using merely two microphones we can separate up to six mixed speech signals under anechoic conditions. The number of source signals is not assumed to be known in advance. It is also possible to maintain the extracted signals as stereo signals.

1. INTRODUCTION

Blind source separation (BSS) 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. The term blind refers to that only the recorded mixtures are known. An important application for BSS is separation of speech signals. The recorded mixtures are assumed to be linear superpositions of the source signals, i.e.

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

where $\mathbf{A}$ is an $M \times N$ mixing matrix and n denotes the discrete time index. $\boldsymbol{\nu}(n)$ is additional noise. A method to retrieve the original signals up to an arbitrary permutation and scaling is independent component analysis (ICA) [1]. In ICA, the main assumption is that the source signals are independent. By applying ICA, an estimate $\mathbf{y}(n)$ of the source signals can be obtained by finding a (pseudo)inverse $\mathbf{W}$ of the mixing matrix so that

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

Many methods require that the number of source signals is known in advance. Another drawback of most of these methods is that the number of source signals is assumed not to exceed the number of microphones, i.e. $M \geq N$. Even if the mixing process $\mathbf{A}$ is known, it is not invertible, and in general, the independent components cannot be recovered exactly [1]. In the case of more sources than sensors, the *overcomplete/underdetermined* case, successful separation often relies on the assumption that the source signals are sparsely distributed - either in the time domain, in the frequency domain or in the time-frequency (T-F) domain [2], [3], [4], [5]. If the source signals do not overlap in the time-frequency domain, high-quality reconstruction could be obtained [4].

However, there is overlap between the source signals. In this case, good separation can still be obtained by applying a binary time-frequency mask to the mixture [3], [4]. In *computational auditory scene analysis*, the technique of T-F masking has been commonly used for years (see e.g. [6]). Here, source separation is based on organizational cues from auditory scene analysis [7]. More recently the technique has also become popular in blind source separation, where separation is based on non-overlapping sources in the T-F domain [8]. T-F masking is applicable to source separation/ segregation using one microphone [6], [9] or more than one microphone [3], [4]. T-F masking can be applied as a binary mask. For a binary mask, each T-F unit is either weighted by one or by zero. In order to reduce musical noise, more smooth masks may also be applied [10]. An advantage of using a binary mask is that only a binary decision has to be made [11]. Such a decision can be based on, e.g., clustering [3], [4], [8], or direction-of-arrival [12]. ICA has been used in different combinations with the binary mask. In [12], separation is performed by removing signals by masking $N - M$ signals and afterwards applying ICA in order to separate the remaining M signals. ICA has also been used the other way around. In [13], it has been 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.

In this paper, a novel method for separating an arbitrary number of speech signals is proposed. Based on the output of a square (2×2) ICA algorithm and binary T-F masks, this method iteratively segregates signals from a mixture until an estimate of each signal is obtained.

2. 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 certain direction and no reflections occur, i.e. an anechoic environment. In order to keep the problem simple, the source signals are mixed by an instantaneous mixing matrix as in eq. (1). Due to delays between the microphones, instantaneous ICA with a real-valued mixing matrix usually is not applica-

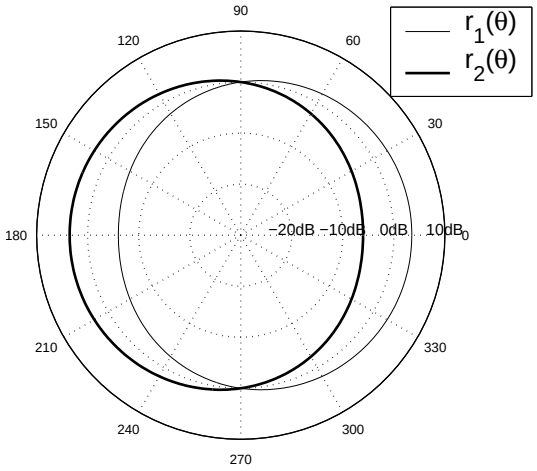

Fig. 1. The two directional microphone responses are shown as function of the direction θ.

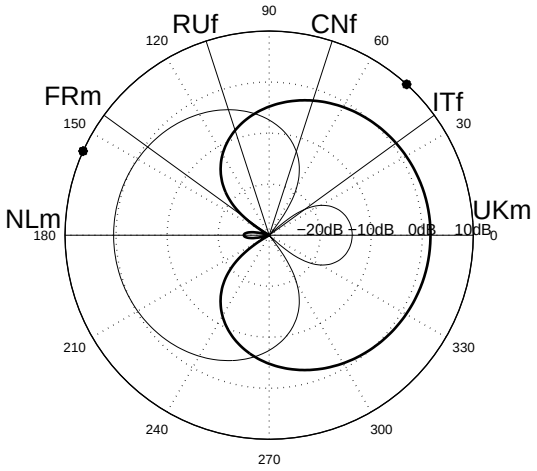

Fig. 2. The polar plots show the gain for different directions. ICA is applied with two sensors and six sources. The two dots at the periphery show the null directions. The lines pointing out from the origin denote the true direction of the speech sources. The three-letter abbreviations (see table 1) identifies the different speech signals which have been used. As it can be seen from the figure, the ICA solution tends to place the null towards sources spatially close to each other. Therefore, each of the two outputs is a group of signals spatially close to each other.

Table 1. The six speech signals. All speakers use raised voice as if they were speaking in a noisy environment.

Abbreviation	Description
CNf	Female speech in Chinese
NLm	Male speech in Dutch
FRm	Male speech in French
ITf	Female speech in Italian
UKm	Male speech in English
RUf	Female speech in Russian

ble to signals recorded at an array of microphones, but if the microphones are placed at exact same location and the microphones have different responses for different directions, the separation of delayed sources can be approximated by the instantaneous model [14]. Hereby, a combination of microphone gains correspond to a certain directional pattern. Therefore, two directional microphone responses are used. The two microphone 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 figure 1. It is possible to make two such directional patterns by adding and subtracting omnidirectional signals from two microphones placed closely together. Hence, the mixing system is given by

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix}. \quad (3)$$

Different speech signals are used as source signals. The used signals are sampled with a sampling frequency of 10 kHz and the duration of each signal is 5 s. The speech signals are shown in table 1.

2.1. More sources than sensors

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 $\mathbf{y}$ is another linear superposition of each of the source signals, where the weights are given by $\mathbf{G} = \mathbf{WA}$ instead of just $\mathbf{A}$ as in (1). Hereby, $\mathbf{G}$ just corresponds to another weighting depending on θ. These weights make $y_1(n)$ and $y_2(n)$ as independent as possible. This is illustrated in figure 2. An implementation of the infomax ICA algorithm [15] has been used. The BGFS method has been used for optimization [16][1]. The figure shows the two estimated spatial responses from $\mathbf{G}(\boldsymbol{\theta})$ in the overdetermined case. The response of the m'th output is given by $|\mathbf{w}_m^T \mathbf{a}(\theta)|$, where $\mathbf{w}_m$ is the separation vector from the m'th output and $\mathbf{a}(\theta)$ is the mixing vector for the arrival direction θ [17]. By varying θ over all possible directions, directivity patterns can be created as shown in figure 2. The estimated null placement is illustrated by the two round dots placed at the periphery of the polar plot. The lines pointing out from the origin illustrate the correct direction of the source signals. Here, the sources are uniformly distributed in the interval $[0^\circ \leq \theta \leq 180^\circ]$. As it can be seen, 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 the first output, $y_1(n)$, the signals NLm and FRm are dominating and in the second output, $y_2(n)$, the signals UKm, ITf and CNf are dominating. The sixth signal, RUf exists in both outputs. This new weighting of the signals can be used to estimate binary masks.

[1]Matlab toolbox available from `http://mole.imm.dtu.dk/toolbox/ica/`

3. BLIND SOURCE EXTRACTION WITH ICA AND BINARY MASKING

A flowchart for the algorithm is given in figure 3. As described in the previous section, a two-input-two-output ICA algorithm is applied to the input mixtures, disregarding the number of source signals that actually exist in the mixture. The two output signals are arbitrarily scaled. The scaling is fixed by using knowledge about the microphone responses. Hereby, the two null directions can be found. The two output signals are scaled such that where one directional response has a null, the other response has a unit gain. The two re-scaled output signals, $\hat{y}_1(n)$ and $\hat{y}_2(n)$ are transformed into the frequency domain e.g. by use of the Short-Time Fourier Transform STFT so that two spectrograms are obtained:

$$\hat{y}_1 \quad \rightarrow \quad Y_1(\omega, t) \tag{4}$$

$$\hat{y}_2 \quad \rightarrow \quad Y_2(\omega, t), \tag{5}$$

where ω denotes the frequency and t is the time index. The binary masks are then determined by for each T-F unit comparing the amplitudes of the two spectrograms:

$$\mathrm{BM1}(\omega, t) \quad = \quad \tau |Y_1(\omega, t)| > |Y_2(\omega, t)| \tag{6}$$

$$\mathrm{BM2}(\omega, t) \quad = \quad \tau |Y_2(\omega, t)| > |Y_1(\omega, t)|, \tag{7}$$

where τ is a threshold. Next, each of the the two binary masks is applied to the original mixtures in the T-F domain, and by this nonlinear processing, some of the speech signals are *removed* by one of the masks while other speakers are removed by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT. If there is only a single signal left in the masked output, defined by the selection criteria in section 3.1, i.e. all but one speech signal have been masked, this signal has been extracted from the mixture and it is saved. If there are more than one signal left in the masked outputs, ICA is applied to the two masked signals again and a new set of masks are created based on (6), (7) and the previous masks. The use of the previous mask ensures that T-F units that have been removed from the mixture are not reintroduced by the next mask. This is done by an element-wise multiplication between the previous mask and the new mask. This iterative procedure is followed until all masked outputs consist of only a single speech signal. Notice, the output signals are maintained as two signals. Stereo signals created with directional microphones placed at the same location with an angle between the directional patterns of 90° (here 180°) are termed XY-stereo.

3.1. Selection criterion

Further processing on a pair of masked signals should be avoided in two cases. If all but one signal have been removed or if too much has been removed so that there is no signal left after applying the mask. The decisions are based on the eigenvalues of the covariance matrix between the masked sensor signals. The covariance matrix is calculated as

$$\mathbf{R} = \langle \hat{\mathbf{x}} \hat{\mathbf{x}}^T \rangle, \tag{8}$$

where $\langle \cdot \rangle$ denotes the expectation with respect to the whole signal, and $\hat{\mathbf{x}}$ is the two time domain signals of which the binary mask has been applied. If $\hat{\mathbf{x}}$ only contains one signal, the covariance matrix is singular, and the smallest eigenvalue $\lambda_{\min}$ is approximately equal to zero [18]. Since parts of the other signals may remain after masking, the smallest eigenvalue is equal to the noise variance

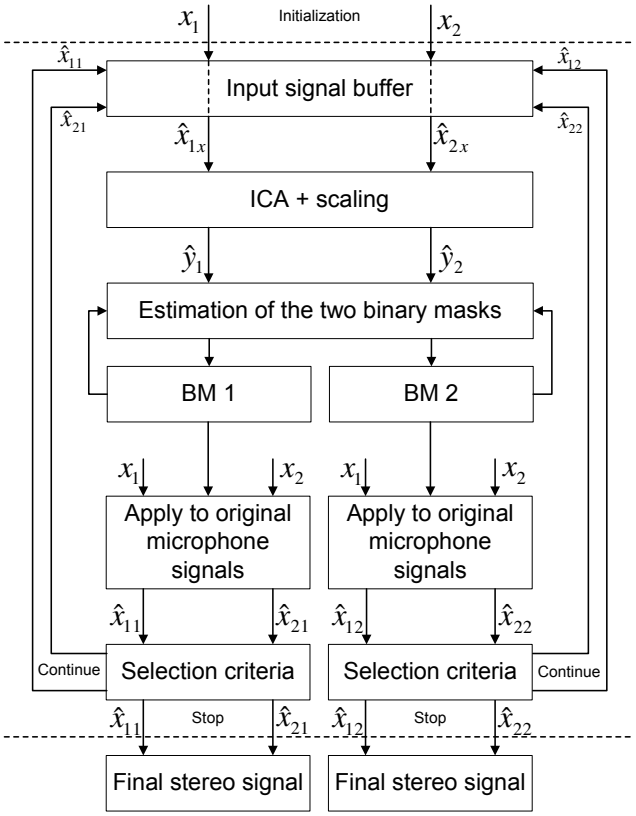

Fig. 3. Flowchart showing the main steps of the proposed algorithm. From the output of the ICA algorithm, binary masks are estimated. The binary masks are applied to the original signals which again are processed through the ICA step. Every time the output from one of the binary masks is detected as a single signal, the signal is stored. The iterative procedure stops when all outputs only consist of a single signal.

of these remaining signals. Therefore, if $\lambda_{\min}$ is smaller than a certain noise threshold $\tau_{\lambda_{\min}}$, it is assumed that there is less than two signals and no further processing is necessary. In order to discriminate between zero or one signal, the largest eigenvalue $\lambda_{\max}$ is considered. If $\lambda_{\max}$ is smaller than a certain threshold $\tau_{\lambda_{\max}}$, the output is considered of such a bad quality that the signal should be thrown away.

3.2. Finding the remaining signals

Since some signals may have been removed by both masks, all T-F units that have not been assigned the value '1' are used to create a *remaining mask*, and the procedure is applied to the mixture signal of which the remaining mask is applied, to ensure that all signals are estimated. Notice, this step has been omitted from figure 3.

4. EVALUATION

The algorithm described above has been implemented and evaluated with mixtures of the six signals from table 1. For the STFT,

an FFT length of 2048 has been used. This gives a frequency resolution of 1025 frequency units. A Hanning window with a length of 512 samples has been applied to the FFT signal and the frame shift is 256 samples. A high frequency resolution is found to be necessary in order to obtain good performance. The sampling frequency of the speech signals is 10 kHz. The three thresholds τ, $\tau_{\lambda_{\min}}$ and $\tau_{\lambda_{\max}}$ have been found from initial experiments. In the ICA step, the separation matrix is initialized by the identity matrix, i.e. $\mathbf{W} = \mathbf{I}$. In order to test robustness, $\mathbf{W}$ was also initialized with a random matrix with values uniformly distributed over the interval [0,1]. The different initialization did not affect the result. When using a binary mask, it is not possible to reconstruct the speech signal as if it was recorded in the absence of the interfering signals, because the signals partly overlap. Therefore, as a computational goal for source separation, the *ideal binary mask* has been suggested [11]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the desired signal to the energy of all the interfering signals. Whenever the signal energy is highest, the T-F unit is assigned the value '1' and whenever the interfering signals have more energy, the T-F unit is assigned the value '0'. As in [9], for each of the separated signals, the percentage of energy loss P_{EL} and the percentage of noise residue P_{NR} are calculated:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \quad (9)$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \quad (10)$$

where $O(n)$ is the estimated signal, and $I(n)$ is the recorded mixture resynthesized 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 $I(n)$. Also the signal to noise ratio (SNR) is found. Here the SNR is defined using the resynthesized speech from the ideal binary mask as the ground truth

$$\text{SNR} = 10\log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n) - O(n))^2}\right]. \quad (11)$$

The algorithm has been applied to mixtures consisting of up to six signals. In all mixing situations, the signals have been uniformly distributed in the interval $[0^\circ \leq \theta \leq 180^\circ]$. The separation results are shown in figure 4 and in table 2.

Two ideal binary masks have been found – one for each microphone signal. In all cases, all the signals have been segregated from the mixture. In most cases also the correct number of signals is estimated. Only in the case of three mixtures, one of the source signals is estimated twice. The double extraction is caused by the selection criteria. Based on the chosen thresholds, the selection criteria in some cases allows a signal to be extracted more than once. In the case of the six mixtures from figure 2, the six estimated binary masks are shown in figure 5 along with the estimated ideal binary masks from each of the two microphone signals. The input SNR (SNR_i) is shown in figure 4 too. The SNR_i is the ratio between the desired signal and the noise in the recorded mixtures. The separation quality decreases when the number of signals is increased. This is expected because when the number of mixed signals is increased, the mixtures become less sparse. Random distributions of the source directions as well as more than six signals have also been examined. Here, in general, not all the sources are separated from each other. If the arrival angles between signals are too narrow, these signals may be detected as a single signal, and they are not separated. Listening tests validate the separation results. This method differs from previous methods which use a binary mask and two microphones [3], [4]. In [3], binaural cues have been applied for separation, i.e interaural time and intensity differences. In [4], the separation is likewise based on amplitude and time difference of each source. Here separation is based on clustering of T-F units that have similar amplitude and phase properties. In our approach too, separation can only be achieved if the source signals have different spatial positions, but the separation criterion is based on independence between the source signals.

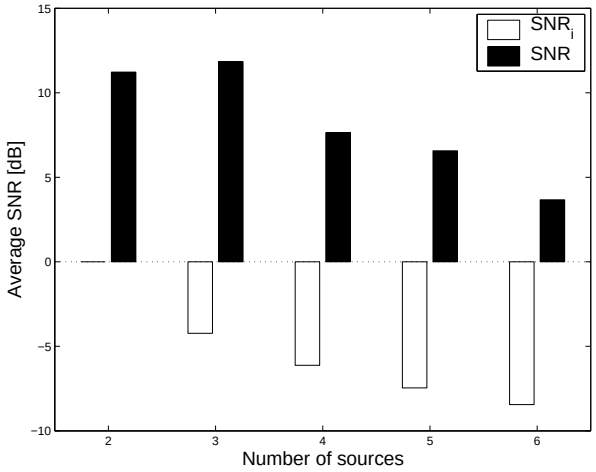

Fig. 4. The signal to noise ratio as function of the number of source signals. The average SNR for the mixtures before separation (SNR_i) is shown as well as the average SNR after separation calculated by eq. (11). In the case of three signals, the incorrectly estimated signal is ignored (see table 2).

5. CONCLUDING REMARKS

A novel method of blind source separation of has been described. Based on sparseness and independence, the method iteratively extracts all the speech signals without knowing the signals in advance. An advantage of this method is that stereo signals are maintained through the processing. So far, the method has been applied to successful separation of up to six speech signals under anechoic conditions by use of two microphones. Future work will include separation of mixtures in reverberant environment, a more blind solution of the scaling problem, and improved techniques for the stopping criteria based on detection of a single signal. Alternative to using a linear frequency scale, a frequency scale that models the auditory system more accurately could be used, because an auditory-based front-end is reported to be more robust than a Fourier-based analysis in the presence of background interference [9]. The use of more than two sensors could also be investigated. By using more than two sensors, a better resolution can be obtained

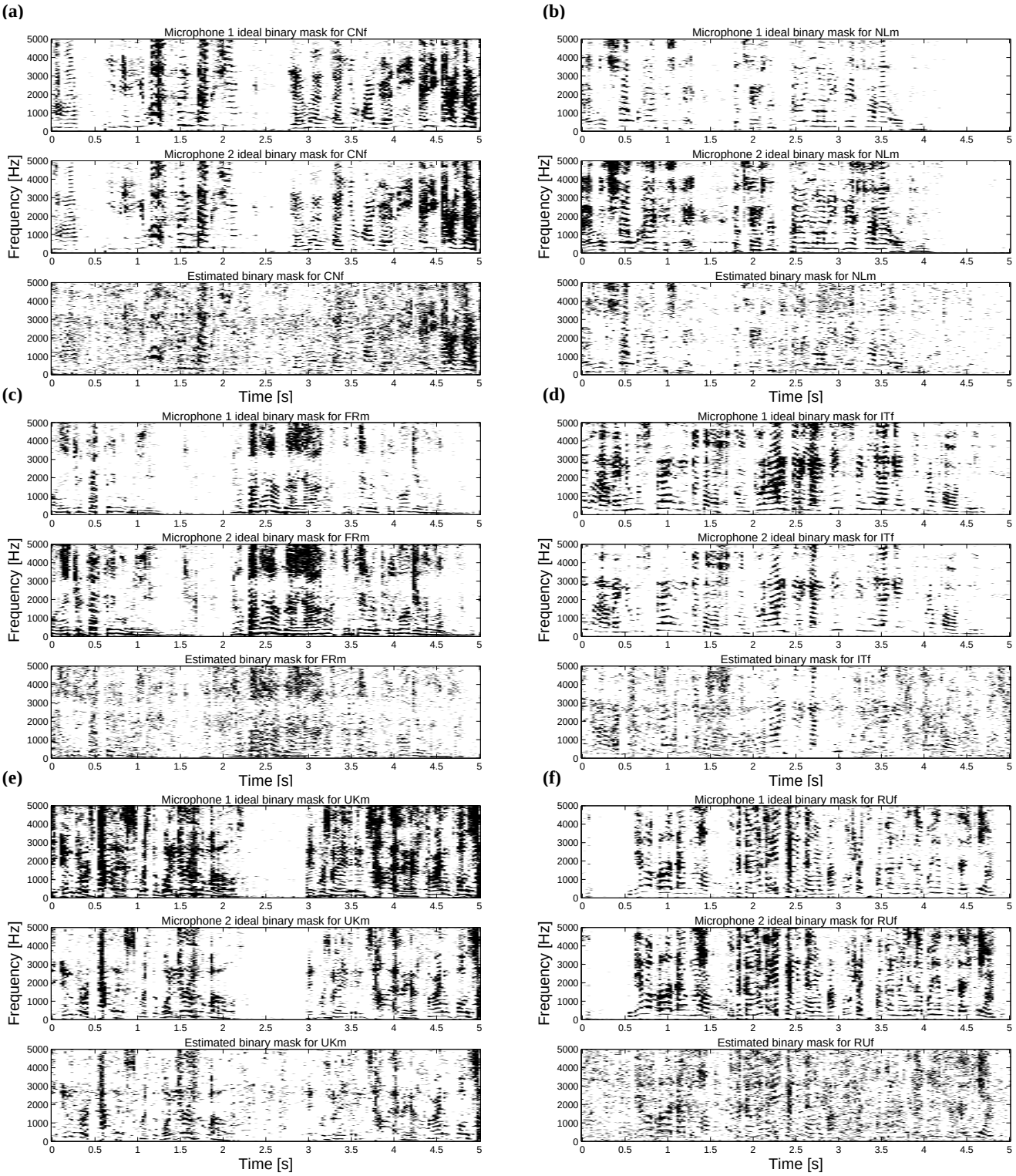

Fig. 5. For a mixture of 6 mixed speech signals, binary masks have been estimated for each of the 6 speech signals. The black areas correspond to the mask value '1' and the white areas correspond to the mask value '0'. The results are shown together with the calculated *ideal binary masks* of each of the two microphone signals. The signals (**a**)–(**f**) appear in the order which they were extracted from the mixture. The first three signals (**a**)–(**c**) were extracted after two iterations, the next two signals (**d**), (**e**) were extracted after three iterations. The last signal (**f**) was extracted from the remaining mask as described in section 3.2.

Table 2. Separation results. Mixtures consisting from two up to six signals have been separated from each other successfully. In most cases, the correct number of sources has been extracted. Only in the case of three source signals, one of the signals has been estimated twice. Here the average performance has been calculated with(†) and without the extra signal. The signals appear in the order which they were extracted from the mixture.

Separated Signal	Microphone 1		Microphone 2	
	$P_{EL}(\%)$	$P_{NR}(\%)$	$P_{EL}(\%)$	$P_{NR}(\%)$
UKm	0.01	8.42	6.83	0.00
FRm	7.13	0.00	0.00	6.11
Average	3.57	4.21	3.41	3.06
NLm	0.11	2.46	3.84	0.06
CNf	5.28	0.16	0.26	2.81
CNf†	86.39	13.12	88.97	63.95
RUf	6.74	11.55	6.17	17.26
Average†	24.63	6.82	24.81	21.02
Average	4.04	4.72	3.43	6.71
CNf	1.27	13.25	3.78	13.79
RUf	2.14	17.64	17.26	3.24
FRm	5.37	2.77	1.01	10.79
UKm	19.60	8.00	14.67	4.60
Average	7.09	10.41	9.18	8.11
RUf	10.65	20.00	24.17	17.70
NLm	8.11	4.13	13.58	1.84
FRm	9.81	17.68	1.32	22.37
ITf	19.20	4.37	4.87	6.92
CNf	4.74	15.55	5.13	16.93
Average	10.50	12.35	9.81	13.15
CNf	8.72	28.20	6.77	21.92
NLm	11.96	15.45	16.32	11.47
FRm	16.05	34.95	29.05	28.72
ITf	29.69	26.87	20.36	23.08
UKm	35.56	6.14	23.26	8.38
RUf	19.58	46.57	28.14	35.33
Average	20.26	26.36	20.65	21.48

and ambiguous arrival angles may be avoided. Also applications for other types of sparse signals could be examined.

6. ACKNOWLEDGEMENTS

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).

7. REFERENCES

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

[2] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time fourier transform," in *Proc. ICA'2000*, 2000, pp. 87–92.

[3] 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.

[4] Ö. 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.

[5] 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.

[6] 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.

[7] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 2 edition, 1990.

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

[9] 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.

[10] 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*, March 18–23 2005, vol. III, pp. 81–84.

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

[12] S. Araki, S. Makino, H. Sawada, and R. Mukai, "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.

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

[14] 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.

[15] 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.

[16] H.B. Nielsen, "Ucminf - an algorithm for unconstrained, nonlinear optimization," Tech. Rep. IMM-TEC-0019, IMM, Technical University of Denmark, 2001.

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

[18] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acous., Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, April 1985.