2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

An Improved permutation solver for blind signal separation based front-ends in robot audition

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano

***Abstract*— The model of the human/machine hands-free speech interface is defined as a point source (the user voice) and a diffuse background noise. This situation is very different from the usual cocktail party model, separation of a mixture of speeches, that is usually treated in frequency domain blind signal separation (FD-BSS). In particular, the fast permutation solvers proposed for the cocktail party model results in poor separation performance in this case. In order to resolve the permutation more efficiently, this paper proposes a new approach that exploits the statistical discrepancy between the target speech and the diffuse background noise.**

I. Introduction

In recent years, the acoustic signal processing community started investigating blind signal separation (BSS) techniques for processing the multidimensional observation given by microphone arrays (see review paper [1]). The frequency domain approach, referred to as FD-BSS, is especially of great interest since the convolutive mixture modeling the reverberant environment can be efficiently processed in the frequency domain. However, a specific problem of this approach is the so called *permutation indeterminacy* that requires the addition of a permutation resolution method to achieve the separation. Most of the research has been focused on the cocktail party problem: the separation of several speech signals [1], [2].

But another problem, that has been overlooked, is the separation of a close target speech signal from a diffuse background noise (created by the sources far from the microphone array). This situation is of great interest since it describes the conditions of the human/machine interaction with a hands-free speech interface. The user interacting with the machine is close to the microphone array whereas other noise sources are at a larger distance. The hands-free speech interface picks the user's voice at distance by a microphone array making a more natural interface with the machine. But the cost is that noises and reverberation deteriorate the speech quality.

In presence of diffuse background noise, the authors in [3] showed that FD-BSS gives a better estimate of the diffuse background noise than of the target speech signal. Consequently they proposed an efficient front-end combining spatial subtraction array techniques with FD-BSS based noise estimation. Unfortunately, the permutation indeterminacy in the presence of diffuse background noise degrades the quality of the noise estimation. In particular, the fast permutation method based on the direction of arrival (DOA) proposed in [4] that is well suited for the cocktail party problem is not a reliable solution in the presence of diffuse background noise.

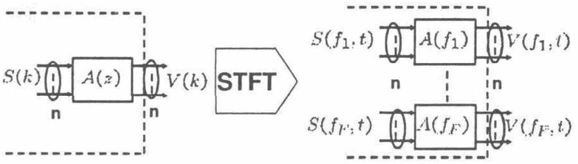

Fig. 1. Time domain convolutive mixture and equivalent frequency domain instantaneous mixtures.

This paper focuses on the permutation resolution for FD-BSS in the situation of the human/machine hands-free speech interface. The main objective is to enhance the quality of the diffuse background noise estimate. The approach presented in this paper exploits the statistical discrepancy between the target speech and the diffuse background noise. This statistical approach specific to the human/machine hands-free speech interface is different of the usual statistical approaches used in FD-BSS (that were usually developed for the cocktail party problem) and does not rely on the same spatial information used by the DOA-based methods. But we also show how to combine the DOA information with the proposed method. Some simulations in a realistic environment are provided to show that the proposed approach significantly improves the quality of the diffuse background noise estimate.

II. Preleminaries

A. Frequency domain blind signal separation

The goal of FD-BSS is to recover some unknown signals when only convolutive mixtures of these signals are observed. Performing the separation in the frequency domain replaces the time domain convolutive mixture by several simpler instantaneous mixtures in the frequency domain (see Fig. 1). The frequency domain model of the mixture is obtain by applying a short time Fourier transform (STFT with a F points analysis frame) to the received signals. The observed signal at the fth frequency bin is

$$V(f,t) = A(f)S(f,t), \tag{1}$$

where the $n \times n$ matrix $A(f)$ represents the instantaneous mixture and $S(f,t) = [s_1(f,t),\ldots,s_n(f,t)]^T$ is the emitted signal at the fth frequency bin (t denotes the frame index and f the frequency bin) .

In the fth frequency bin, the estimates $Y(f,t) = [y_1(f,t),\ldots,y_n(f,t)]^T$ are obtained by applying an unmixing matrices $B(f)$ to the observed signals (see Fig.2)

$$Y(f,t) = B(f)V(f,t) = B(f)A(f)S(f,t). \tag{2}$$

Graduate school of information science Nara Institute of science and technology Ikoma, Nara, Japan even@is.naist.jp

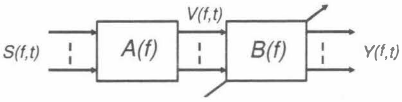

Fig. 2. Mixture and blind separation at frequency bin f.

The blind separation is possible because of the following theorem [5]

Theorem 1

If the components of $S(f,t)$ are statistically independent then the component of $Y(f,t)$ are statistically independent if and only if $B(f)$ is such that

$$Y(f,t) = P(f)\Lambda(f)S(f,t)$$

where $P(f)$ is a $n \times n$ permutation matrix and $\Lambda(f)$ is a diagonal $n \times n$ matrix.□

As a consequence, in each frequency bin, it is possible to recover the components of $S(f,t)$ up to scale and permutation indeterminacy by finding the unmixing matrix $B(f)$ that gives an estimate with statistically independent components (see review paper [1]).

Theorem 1 also shows one major difficulty of the FD-BSS approach: The *permutation indeterminacy*. Considering the first component of $Y(f,t)$ after separation of the frequency domain signals in all the bins we have

$$\begin{bmatrix} y_1(1,t) \\ \vdots \\ y_1(F,t) \end{bmatrix} = \begin{bmatrix} P(1)^{(1,:)}S(1,t) \\ \vdots \\ P(F)^{(1,:)}S(F,t) \end{bmatrix},$$

where $P(f)^{(1,:)}$ denotes the first row of $P(f)$ (for simplicity $\Lambda(f)$ is omitted; see end of Sect. III-B for scale indeterminacy).

Transforming back $[y_1(1,t),\ldots,y_1(F,t)]^T$ to the time domain only gives a time estimate $y_1(k)$ of one of the original time signals (say $s_i(k)$) if $y_1(f,t) = s_i(f,t)$ for all f. Namely the components belonging to the same signal must be matched across all the frequency bins before transforming back the signals to the time domain. The methods used to force all the $P(f)$ to be equal are referred to as *permutation resolution methods*.

Several general permutation resolution methods exploit statistical dependency between the components $y_i(f,t)$. These methods compute second order [6] or higher order statistics [7] between components in different bins and match corresponding components using this statistical information. The main drawback of these methods is their computation cost. For this reason other approaches are usually preferred like the method in [8]. But this method is not well suited for long filters [1].

B. Point source separation: the two speakers scenario

The research in FD-BSS for acoustic signal processing mainly focused on speech/speech separation also known as the cocktail party problem. As a result the conventional permutation resolution methods are designed with the cocktail party model in mind. In particular, a speech signal can be approximately considered as a point source and its direction of arrival (DOA) can be exploited to resolve permutation. The DOA of the speech is the direction of the emission point as seen from the microphone array.

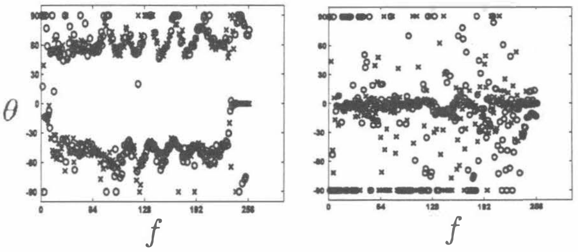

Fig. 3. DOA estimate versus frequency bins speech/speech (left) and speech/diffuse noise (right).

In [2], the authors showed that FD-BSS is equivalent to a set of adaptive null beamformers (ANBF) each having its null toward different speakers. Then it is possible to resolve the permutation by determining the position of these nulls using the directivity pattern of the matrices $B(f)$.

In [4], a method exploiting the DOA information that does not requires the estimation of the directivity pattern is proposed.

Considering two persons talking at the same time in a room with a reverberation time of 200ms. Suppose the speakers are distant of 1 meter from a two microphone array and have DOAs of $\theta_1 = 60$ and $\theta_2 = -60$. Applying the method in [4], the estimated DOA of the first components $y_1(f,t)$ ($\circ$) and of the second components $y_2(f,t)$ ($\times$) are plotted in Fig. 3(left). The DOA repartition in the different bins shows two clusters around θ_1 and θ_2. Then it is easy to determine which components to match together. (Note that even if the speakers are close to the microphone array and the reverberation time is small, the variance of the DOA estimate is large due to the reverberation condition).

III. Main results

A. Target speech in diffuse background noise

The model of the human/robot hands-free speech interface is very different from the cocktail party model. The user is assumed to be close to the microphone array and thus is modeled as a point source. But the other sources are far from the microphone array thus because of the reverberant environment they are seen as a diffuse background noise.

In particular, the speech has a well defined DOA (few spreading) but the diffuse background noise has no clear DOA. Consequently, using FD-BSS it is possible to place a null in the direction of the speech and get a good estimate of the diffuse noise. But it is not possible to get a good speech estimate since with a limited number of microphones it is not possible to cancel the diffuse background noise. This is the reason why FD-BSS gives a good estimate of the diffuse background noise but not of the target speech [3].

The fact that the diffuse background noise has no clear DOA also prevent the use of the fast DOA based permutation resolution method [4]. Fig. 3 (right) shows the DOA estimate corresponding to the human/robot hands-free speech

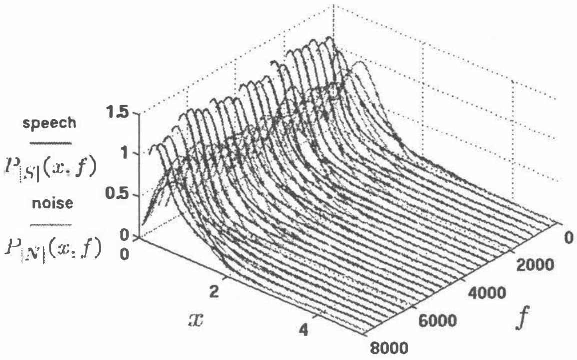

Fig. 4. Pdf of speech and noise versus frequency bins.

interface situation (corresponding to the data in Sect. IV). In some frequency bins, it is not possible to determine which estimated DOA correspond to the speech or to the diffuse background noise. In such frequency bins, the DOA based permutation solver is likely to wrongly select the target speech as diffuse background noise component. Then the wrongly selected target speech component deteriorates the diffuse background noise estimate.

Note that the elegant method proposed in [9] that is able to separate signal with no permutation requires to have at least as many sensors than there are signals which is not possible for the diffuse background noise (composed of many signals).

B. Statistical discrepancy of speech and diffuse noise

The spatial information contained in the separation matrices $B(f)$ cannot be exploited as it is when considering the human/robot hands-free speech interface. Thus we propose to use the statistical information contained in the estimates $y_i(f,t)$. But contrary to the general methods cited at the end of Sect.II-A, we do not compute statistics involving components from different frequency bins. In the case of a target speech in a diffuse background noise, using statistics computed in each frequency bins is enough to resolve the permutation.

In time domain, the distribution of the speech signal amplitude is often modeled by a Laplacian distribution because the speech is a non stationary signal having activity and non activity parts (silence). On the contrary, the diffuse background noise is composed of the superposition of many sounds consequently its amplitude has a distribution that is close to the Gaussian distribution. After the STFT, in each of the frequency bins, we can also observe that the modulus of the speech signal has a spikier distribution than that of the diffuse background noise. Fig. 4 shows the pdf $P_{|S|}(x,f)$ of the modulus of the normalized speech and the pdf $P_{|N|}(x,f)$ of the modulus of the normalized noise for a speech and a diffuse background noise recorded in a train station (see Sect. IV). This statistical discrepancy between speech and diffuse background noise is the key of the proposed method. In each frequency bins, after convergence of the separation matrices $B(f)$, the permutation resolution is performed using statistical features computed on the components of $Y(f,t)$. Then the selection of the components corresponding to the target speech or diffuse background noise is based on this feature. Several different features can be derived from this idea.

In the following section we present different exploitation of this idea. But let us briefly define the diffuse background noise estimate in more detail before presenting these methods.

One important property is that to obtain the estimation of the diffuse background noise contribution at the microphone array only the selection of the speech component is required. Suppose that at frequency bin f the estimated speech component is $y_i(f,t)$ then the projection back of the diffuse background noise is

$$Z(f,t) \;=\; B(f)^{-1}(I-D_i)Y(f,t)$$

where I is the identity matrix and D_i is a matrix having only one non null entry $d_{ii} = 1$. If we assume perfect separation $B(f)A(f) = P(f)\Lambda(f)$ and $s_1(f,t)$ is the speech component then $P(f)$ is such that $P(f)^{-1}D_iP(f) = D_1$ and

$$\begin{aligned} Z(f,t) &= A(f)S(f,t) - A(f)\Lambda(f)^{-1}D_1\Lambda(f)S(f,t) \\ &= A(f)S(f,t) - A(f)D_1S(f,t) \\ &= A(f)S(f,t) - A(f,t)^{(:,1)}s_1(f,t) \\ &= \sum_{k=2}^{n} A(f,t)^{(:,k)}s_k(f,t) \end{aligned}$$

where $A(f,t)^{(:,j)}$ is the j^{th} column of $A(f,t)$. Namely Z(f,t) is equal to the contribution of the diffuse background noise at the microphone array. Note that the scale indeterminacy $\Lambda(f)$ is compensated in the process of projection back.

C. Proposed permutation resolution methods

1) Kurtosis based method: The statistical feature computed for all the separated components is the kurtosis of their modulus

$$K_{y_i,f} = \frac{\mathcal{E}\left\{|y_i(f,t)|^4\right\}}{\mathcal{E}\left\{|y_i(f,t)|^2\right\}^2} - 3.$$

In each frequency bin, the component with the largest kurtosis is selected as the speech component.

2) Distribution fitting based method: The statistical feature computed for all the separated components is the scale parameter $\alpha_{y_i,f}$ of the Laplacian distribution that fits the pdf of the modulus. The maximum likelihood estimate of this parameter is

$$\alpha_{y_i,f} = \frac{1}{\mathcal{E}\left\{|y_i(f,t)|\right\}}.$$

In each frequency bin, the component with the largest parameter is selected as the speech component.

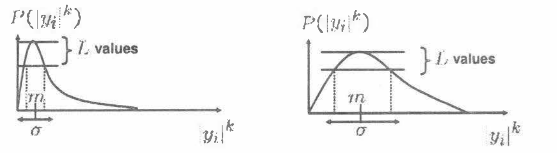

Fig. 5. Distributions of modulus for target speech (left) and diffuse noise (right).

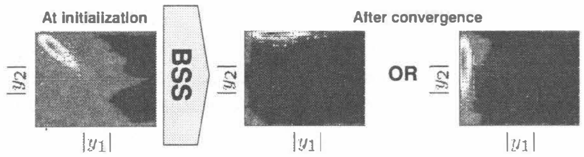

Fig. 6. Joint pdf of observation (left) and estimates (center) (right) at frequency bin f.

3) Pdf based method: The statistical feature is obtained from a kernel based estimate of the pdf of the modulus $|y_i(f,t)|$ (For some separation methods using adaptive method to determine the $B(f)$ such estimates of the pdf are readily available after separation [10]). The estimate of the distribution obtained from $|y_i(f,t)|$ for the frames $t \in [1:T]$ is a set of $K < T$ couples

$$\{|y_i|^{(k)}, P^{(k)}_{|y_i|}\}$$

with $k \in [1:K]$. The values are obtained by convoluting a K bins histogram of the data with a Gaussian Kernel (adapted from [11]). To determine the discriminant statistical feature that measures the spikiness of the distribution, we first select the L couples

$$\{|y_i|^{(k_l)}, P^{(k_l)}_{|y_i|}\}_{l\in[1:L]}$$

with highest values of $P^{(k)}_{|y_i|}$ (see Fig. 5). Then we compute the mean $\overline{m_i}$ and standard deviation σ_i of the L selected values of the amplitude $|y_i|^{(k_l)}_{l\in[1:L]}$. Finally the statistical feature is obtained by taking

$$I(y_i(f,t)) = \overline{m_i}\ \sigma_i.$$

At the frequency bin f, after computing the statistical feature for all the components of $Y(f,t)$, the component with the smallest feature is selected as the target speech.

4) Joint pdf based method: For the two microphone array case, it is possible to use the joint pdf of the estimated components to resolve the permutation. Fig. 6 (left) shows such joint pdf $P(|y_1|,|y_2|)$ at frequency bin f before the separation is performed. After separation by the FD-BSS method, the joint pdf take the form in Fig. 6 (center) if $y_1(f,t)$ estimates the speech component or in Fig. 6 (right) if $y_2(f,t)$ estimates the speech component. The classification of the joint pdf in the one in Fig. 6 (center) or Fig. 6 (right) is performed by comparing the average values of $|y_1|$ and $|y_2|$ computed for the L points of maximal density (the joint pdf is estimated on a $K \times K$ grid by convoluting a $K \times K$ 2D histogram of the data with a 2D Gaussian Kernel [11]).

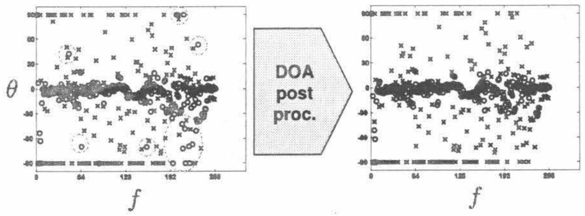

Fig. 7. Using DOA estimate as post processing.

5) Post processing using DOA: It is possible to improve the performance of the proposed methods by using the DOA estimate in a post processing stage.

After applying one of the proposed methods, the DOA is estimated from the permuted $B(f)$ using the estimator from [4]. Fig. 7 (left) shows the DOA estimates obtained after the pdf based method using the same data as in Fig. 3 (right). At some frequency bins, the DOA estimates corresponding to the estimated speech $\circ$ is away from the general trend, see the circled areas in Fig. 7 (left). As the DOA of the speech signal should be consistent along the frequency axis, the row of the separation matrix are permuted in these bins, see Fig. 7 (right). This post processing using the DOA information enables the detection of bins where the statistical index wrongly permuted the speech and noise. These bins appear when the diffuse background noise has a spikier distribution than that of the speech because of an isolated event with high energy (in this data set Sect. IV, it seem to correspond to a bell ringing for announcing incoming train in the station).

IV. EXPERIMENTAL RESULTS

Some experiments were conducted using recording from a train station hall (see Fig.9). A two microphone array (mic. spacing: 2.15cm) was used to record the ambient diffuse noise in the station hall. The impulse response from a speaker at 50cm and 150cm in front of the array were also measured (the reverberation time is RT60 $\approx$ 1s).

In such situation FD-BSS cannot estimate the speech as it cannot suppress the background noise. But FD-BSS can be used to estimate the background noise as proposed in [3] where the noise estimate is used to clean the speech by spectral subtraction. Here, the goal of the simulation is to show that the proposed permutation resolution improves the quality of the diffuse background noise estimate compared to the conventional DOA based method.

The test data is composed of 200 Japanese sentences (JNAS database [12]) that are convoluted with the measured impulse response and mixed with the recorded noise at different SNR levels. The sampling frequency is 16khz and the sentences have variable lengths (from 2.4s to 14.7s). The STFT is performed with a 512 points hanning window with 256 points overlap.

The ICA algorithm is a modified INFOMAX algorithm [13] where the nonlinear activation functions are estimated from the data using a kernel based estimator [10]. The matrices $B(f)$ are initialized to identity in all frequency bins then 300 iterations are performed with an initial adaptation step of $\mu_0 = 0.3$ that is divided by two every 100 iterations.

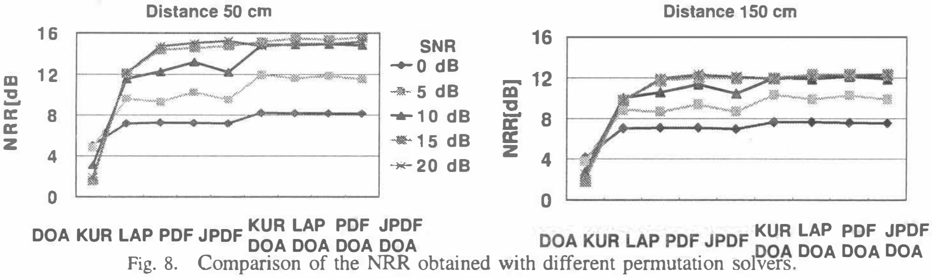

Fig. 8. Comparison of the NRR obtained with different permutation solvers.

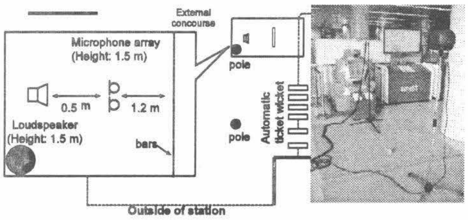

Fig. 9. Experimental settings.

TABLE I
MEAN COMPUTATION TIMES (MS)

	no DOA post proc.	DOA post proc.
DOA	160	(not applicable)
KURT	62	316
LAP	80	333
PDF	352	604
JPDF	2093	2359

In all experiments we compared the approach from [4] (referred to as DOA) to the proposed approaches. KUR, LAP, PDF and JPDF denote respectively the kurtosis based method, the Laplacian distribution fitting based method, the pdf based method and the method exploiting the joint pdf. The terms KUR DOA, LAP DOA, PDF DOA and JPDF DOA refer to the previous methods combined with the proposed DOA post processing. For the PDF method the parameters are $K = 300$ and $L = 10$ (equivalent parameters for JPDF are $K = 50$ and $L = 20$).

The diffuse background noise estimation quality is measure in term of noise reduction rate (NRR) [2] defined as the difference of the SNR of the diffuse background noise estimates (after processing) and the SNR of the observations (before processing). While computing the SNR, the signal of interest is the diffuse background noise and the target speech is considered as the noise. Consequently, a positive NRR means that the diffuse background noise estimate quality is improved since it contains less target speech.

We can see that the proposed methods outperforms the DOA-based method for all situations in Figs.8. The difference is particularly important for higher SNR where any permuted speech component has a high energy that considerably degrades the quality of the estimation. Note that for 150cm the stronger reverberation results in a less good noise estimate because the point source approximation for the target speech is less valid, consequently the spatial null steered by FD-BSS in the direction of the speech is less efficient at canceling it from the noise estimate.
The DOA post processing especially improves the NRR for lower SNRs (0 dB, 5 dB and 10 dB) and the improvement is reduced when the distance increases (one reason is that the variance of the DOA estimate is larger).

The average computational times of the different approaches are given in Table I (in ms). The kurtosis based method and distribution fitting method are faster than the DOA based method in [4] (in fact the distribution fitting requires less computation but it's code was not as optimized as the code of the kurtosis method). The pdf based method requires more computation because of the pdf estimation. The joint pdf approach has the highest cost because of the costly estimation of the joint pdf (but in some situations the additional information contained in the joint density results in better performance). The additional cost of the DOA post processing is comparable to the cost of the DOA or pdf based method. As a result the best methods in tern of performance/cost are the kurtosis based, distribution fitting based and pdf based methods with DOA post processing (at higher SNR the DOA post processing may be discarded). As a comparison, the mean computation time of the FD-BSS separation was 77098 ms (quite long because of the 300 iterations). Thus the ratio of the permutation method computation time to total computation time varies from 0.08% for kurtosis based method to 2.96% for joint pdf based method with DOA post processing (for 100 iterations this ratio would varies from 0.24% to 8.4%).

V. Conclusion

In this paper, we proposed a new approach to the permutation problem in order to improve the performance of FD-BSS based front-end in the presence of diffuse background noise. This approach is more efficient than the traditional DOA-based method as it significantly improves the diffuse background noise estimate quality. As a future development, we are working on combining the proposed method with the DOA approach to improve simultaneous speakers separation in diffuse background noise (equivalent to two concurrent users of the machine).

References

[1] M.S. Pedersen et al., "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Communication*, 2007.

[2] H. Saruwatari et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP Jour. on Appl. Sig. Proc.*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[3] Y. Takahashi et al., "Blind spatial subtraction array with independent component analysys for hands-free speech recognition," *IWAENC (CD-ROM)*, 2006.

[4] H. Sawada et al., "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Proc.*, vol. 12, pp. 530-538, 2004.

[5] P. Comon, "Independent component analysis, a new concept ?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[6] V. Capdevielle, C. Serviere, and J.L. Lacoume, "Blind separation of wide-band sources in the frequency domain," *Proc.ICASSP95*, pp. 2080-2083, 1995.

[7] C. Mejuto, A. Dapena, and L. Castedo, "Frequency-. domain infomax for blind separation of convolutive mixtures," *Proc. ICA 2000*, pp. 315–320, 2000.

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

[9] T. Kim et al., "Independent vector analysis: An extension of ica to multivariates components," *ICA'06*, pp. 165–172, 2006.

[10] N. Vlassis and Y. Motomura, "Efficient source adaptivity in independent component analysis.," *IEEE Trans. Neural Networks*, vol. 12, no. 3, pp. 559–566, 2001.

[11] B.W. Silverman, "Kernel density estimation using the fast fourier transform," *J. Roy. Statist. Soc. Ser. C: Appl. Statist.*, vol. 31, no. 1, pp. 93-99, 1982.

[12] K. Ito et al., "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoustical Society of Japan*, vol. 20, pp. 196-206, 1999.

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