BLIND SEPARATION OF SINGLE CHANNEL MIXTURE USING ICA BASIS FUNCTIONS

Gil-Jin Jang[1,2], *Te-Won Lee*[1], *and Yung-Hwan Oh*[2]

[1]Institute for Neural Computation, University of California, San Diego
La Jolla, California 92093, USA

[2]Spoken Language Laboratory, Department of Computer Science
Korea Advanced Institute of Science and Technology
Taejon 305-701, Korea

{jangbal,tewon}@inc.ucsd.edu, yhoh@cs.kaist.ac.kr
http://inc.ucsd.edu/~jangbal

ABSTRACT

A new technique has been developed to enable blind source separation given only a single channel recording. The proposed method infers source signals and their contribution factors at each time point by a number of adaptation steps maximizing log-likelihood of the estimated source parameters given the observed single channel data and sets of basis functions. This inferencing is possible due to the prior information on the inherent time structure of the sound sources by learning *a priori* sets of time-domain basis functions and the associated coefficient densities that encode the sources in a statistically efficient manner. A flexible model for density estimation allows accurate modeling of the observation and our experimental results show close-to-perfect separation on simulated mixtures as well as recordings in a real environment employing mixtures of two different sources.

1. INTRODUCTION

The need for extracting individual sound sources from mixtures of different signals is increasing in both the commercial and scientific fields. The problem is formulated as: it is assumed that the observed time series $\mathbf{Y} = [y(1)\ y(2)\ \ldots\ y(T)]$ is an addition of M independent sources

$$\mathbf{Y} = \lambda_1 \mathbf{X}_1 + \lambda_2 \mathbf{X}_2 + \ldots + \lambda_M \mathbf{X}_M\,, \qquad (1)$$

where $\mathbf{X}_i$ is the time series of the i^{th} source, and the positive constant λ_i determines the degree of participation or the realized scale of each source to the observations. The goal is to recover the original time series $\mathbf{X}_i$ given only observed single channel input $\mathbf{Y}$. It is convenient assuming $\mathbf{X}_i$ to have zero mean and unit variance. The mixing is non-stationary, in a sense that all the sources are unknown and time-varying.

Various sophisticated methods have been proposed in the research areas such as computational auditory scene analysis (CASA) [1, 2] and independent component analysis (ICA) [3]. Separation algorithms in CASA are based on isolating auditory streams in time or frequency domain by assuming the sparseness of the sources, that is, the observed instances of the individual sources are mutually exclusive in time samples or in spectral domain. Previous work tried to localize the acoustic objects into separate streams, such as classifying speech segments into the same pitch (F_0) groups [4] or decorrelating frequency bands [5]. Recently Roweis [6] has presented a refiltering technique which estimates λ_i in equation 1 as time-varying masking filters that localize sound streams in powerspectral domain. In his work sound sources are supposedly disjoint in the spectrogram and there exists a "mask" that divides completely multiple streams. These approaches are however able to be applied to certain limited environments due to the intuitive prior knowledge of the sources such as harmonic modulations or temporal coherency of the acoustic objects.

The ICA algorithms are data driven methods and relax the strong frequency characteristical assumptions. However ICA algorithms perform best only able when the number of the observed signals are greater or equal the number of sources [3]. Although some recent overcomplete representation may relax this assumption the problem of separating sources from a single channel observation remains difficult. In other aspects, ICA has

been shown highly effective in encoding patterns, including images [7] and natural sounds [8]. The basis functions and the coefficients learned by ICA reflect the statistical structures of the sources, by estimating the maximum likelihood densities.

This paper introduces a technique for single channel blind source separation utilizing the ICA basis functions. The algorithm recovers original sound streams in a number of gradient-ascent adaptation steps maximizing the log-likelihood of the separated signals, which is calculated by the likelihood of their associated coefficients for the given basis functions. The densities of the source coefficients are modeled by generalized Gaussian priors [9] that estimate wide range of probability density functions. The experimental results showed that two different sources were almost perfectly recovered in the simulated mixtures of rock and jazz music, and male and female speech signals, as well as in the real recordings of mixed speech signals and music sound.

2. ADAPTING ICA BASIS FUNCTIONS AND MODEL PARAMETERS

The proposed method first involves the learning of the time-domain basis functions of the sound sources that we are interested in separating. This corresponds to the prior information necessary to successfully separate the signals.

In the case of sound sources, ICA assumes that a segment sampled from a contiguous signal is constructed by a linear superposition of basis functions with scalar multiples. This technique was employed in [10] to learn the basis of speech signals. For the segment of size N starting at time t, $\mathbf{x}(t) = [x(t)\ x(t+1)\ \ldots\ x(t+N-1)]^T$, ICA assumes an unknown source vector $\mathbf{s}(t)$ that are statistically independent. The sources are not observed directly but as a linear combination such that

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{N} \mathbf{a}_i s_i(t), \qquad (2)$$

where $\mathbf{A}$ is a $N \times N$ square matrix of real elements. The columns of $\mathbf{A}$, $\{\mathbf{a}_i\}$, are called the basis functions generating the segments of the observed signal in the real world whereas $\mathbf{W} = \mathbf{A}^{-1}$ refers to the ICA filters that transform the segments into activations or source coefficients $\mathbf{s}(t) = \mathbf{W}\mathbf{x}(t)$.

The ICA learning algorithm is equivalent to maximizing the densities of the corresponding source vectors for the given training data, as well as searching for the linear transformation that make the components as statistically independent as possible [11],

$$\begin{aligned} \mathbf{W}^* &= \arg\max_{\mathbf{W}} \prod_t P\left(\mathbf{x}(t)|\mathbf{W}\right) \\ &= \arg\max_{\mathbf{W}} \prod_t \left\{ \prod_i P\left(s_i(t)\right) \right\} |\det \mathbf{W}|, \quad (3) \end{aligned}$$

Independency between the components factorizes the joint probability densities of the coefficients into the product of marginal ones, and independency over time does on the segments. What matters is therefore how well approximated the model distribution is to the true underlying distribution of $p(s_i)$. We use a flexible prior known as generalized Gaussian or exponential power distribution which models density functions that are peaky and symmetric at the mean, with the varying degree of normality in the following general form [9]:

$$p(s|\mu,\sigma,q) = \frac{\omega(q)}{\sigma} \exp\left[-c(q)\left|\frac{s-\mu}{\sigma}\right|^q\right], \qquad (4)$$

where $\mu = E[s]$, $\sigma = \sqrt{E[(s-\mu)^2]}$, $c(q) = \left[\frac{\Gamma[3/q]}{\Gamma[1/q]}\right]^{q/2}$, and $\omega(q) = \frac{\Gamma[3/q]^{1/2}}{(2/q)\Gamma[1/q]^{3/2}}$. The exponent q controls the distribution's deviation from normality. The Gaussian, Laplacian, and strong Laplacian —speech signals— distributions are modeled by putting $q = 2$, $q = 1$, and $q < 1$ respectively.

3. SOURCE SEPARATION OF SINGLE CHANNEL OBSERVATION

Given the generalized Gaussian model parameters we perform log-likelihood maximization on the source signals to estimate the original sources. Scaling factors of the generative model are learned as well.

3.1. Deriving Learning Rules for Source Signals

The goal of the algorithm is to infer multiple sources from a single mixture given the basis functions in time domain. We consider $M = 2$ only in equation 1. For an observed mixture series $\mathbf{Y}$, we assume that $\mathbf{Y} = \mathbf{Y}_1 + \mathbf{Y}_2$ which are characterized by sets of parameters $\{\lambda_1, \mathbf{W}_1\}$ and $\{\lambda_2, \mathbf{W}_2\}$, where $\mathbf{W}_i$ is the basis filter matrix of $\mathbf{X}_i$, and $\mathbf{Y}_i = \lambda_i \mathbf{X}_i$. To simplify the inferring steps, we force the sum of the factors to be constant: e.g. $\lambda_1 + \lambda_2 = 1$. The likelihood of $\mathbf{Y}_1$ given the observation is replaced by the multiplication of the marginal ones of $\mathbf{Y}_1$ and $\mathbf{Y}_2$:

$$\begin{aligned} &P(\mathbf{Y}_1|\mathbf{Y}, \lambda_1, \mathbf{W}_1, \lambda_2, \mathbf{W}_2) \\ &\quad = P(\mathbf{Y}_1|\lambda_1, \mathbf{W}_1) P(\mathbf{Y}_2|\lambda_2, \mathbf{W}_2) \\ &\quad = P(\mathbf{Y}_1|\lambda_1, \mathbf{W}_1) P(\mathbf{Y} - \mathbf{Y}_1|1-\lambda_1, \mathbf{W}_2)\,. \quad (5) \end{aligned}$$

At every time point $t \in [1, T-N+1]$ a segment $\mathbf{y}_1(t)$ of contiguous N samples is extracted from $\mathbf{Y}_1$. The basis filter matrix $\mathbf{W}_1$ then infers the independent source vector $\mathbf{s}_1(t) = \frac{1}{\lambda_1}\mathbf{W}_1\mathbf{y}_1(t)$ by substituting $\mathbf{x}(t)$ in equation 2 with $\frac{1}{\lambda_1}\mathbf{y}_1(t)$. Respectively $\mathbf{W}_2$ infers $\mathbf{s}_2(t) = \frac{1}{\lambda_2}\mathbf{W}_2\mathbf{y}_2(t)$ where $\mathbf{y}_2(t)$ is from $\mathbf{Y}_2$. We define the likelihood of $\mathbf{Y}_1$ at time t by that of $\mathbf{y}_1(t)$:

$$P(\mathbf{y}_1(t)|\lambda_1, \mathbf{W}_1) = p\left(\mathbf{s}_1(t)\right) |\det \mathbf{W}_1| , \qquad (6)$$

where $p(\cdot)$ is the exponential power distribution of corresponding source. Assuming the independency over time, the conditional probability of $\mathbf{Y}_1$ is obtained from the marginal ones of all the segments,

$$\begin{aligned} P(\mathbf{Y}_1|\lambda_1, \mathbf{W}_1) &= \prod_{t=1}^{T_N} P(\mathbf{y}_{1t}|\lambda_1, \mathbf{W}_1) \\ &= \prod_{t=1}^{T_N} p(\mathbf{s}_{1t}) |\det \mathbf{W}_1| , \end{aligned} \qquad (7)$$

where, for convenience, $T_N = T - N + 1$ and time has become subscripted in vectors and their components throughout the rest of the manuscript. We denote the log of equation 5 by $\mathcal{L}$ and calculate it using equation 7 as

$$\begin{aligned} \mathcal{L} &= \log P(\mathbf{Y}_1|\lambda_1, \mathbf{W}_1) P(\mathbf{Y}_2|\lambda_2, \mathbf{W}_2) \\ &= \sum_{t=1}^{T_N} \left[\log p\left(\mathbf{s}_{1t}\right) + \log p\left(\mathbf{s}_{2t}\right)\right] \\ &\quad + T_N \log |\det \mathbf{W}_1| |\det \mathbf{W}_2| . \end{aligned} \qquad (8)$$

Our interest is in adapting $\mathbf{Y}_1$, i.e. $y_1(t)$ for $\forall t \in [1, T]$, toward the maximum of the objective function $\mathcal{L}$. We derive a gradient-ascent learning rule for $y_1(t)$ by summing up the gradients over all the speech segments where the sample lies:

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial y_1(t)} \\ &= \sum_{n=1}^{N} \left[\frac{\partial}{\partial y_1(t)} \log p(\mathbf{s}_{1t_n}) + \frac{\partial}{\partial y_1(t)} \log p(\mathbf{s}_{2t_n}) \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{N} \left[\varphi(s_{1kt_n}) \frac{w_{1kn}}{\lambda_1} - \varphi(s_{2kt_n}) \frac{w_{2kn}}{\lambda_2} \right] \\ &\propto \sum_{n=1}^{N} \sum_{k=1}^{N} \left[\lambda_2 \varphi(s_{1kt_n}) w_{1kn} - \lambda_1 \varphi(s_{2kt_n}) w_{2kn} \right] \end{aligned} \qquad (9)$$

which is derived by the fact that

$$\frac{\partial s_{kt_n}}{\partial y(t)} = \frac{\partial (\mathbf{w}_k \mathbf{y}_{t_n} / \lambda)}{\partial y(t)} = \frac{w_{kn}}{\lambda} \qquad (10)$$

and

$$\frac{\partial y_2}{\partial y_1} = \frac{\partial (y - y_1)}{\partial y_1} = -1 , \qquad (11)$$

where $t_n = t - n + 1$, $w_{ikn} = \mathbf{W}_i(k, n)$, and

$$\varphi(s) = \frac{\partial \log p(s)}{\partial s} = -\frac{cq}{\sigma^q} |s|^{q-1} \mathrm{sign}(s) , \qquad (12)$$

expressed by the parameters q, c, and σ of the generalized Gaussian for $p(s)$ as defined in equation 4. It is assumed that the mean of source signal s is zero so μ has been eliminated. The final learning rule follows more simply

$$\Delta y_1(t) \propto \sum_{n=1}^{N} \left[\lambda_2 \varphi(\mathbf{s}_{1t_n})^T \mathbf{w}_{1(\cdot,n)} - \lambda_1 \varphi(\mathbf{s}_{2t_n})^T \mathbf{w}_{2(\cdot,n)} \right] , \qquad (13)$$

where $\mathbf{w}_{i(\cdot,n)}$ is the n^{th} column vector of $\mathbf{W}_i$. Because $\partial\mathcal{L}/\partial y_2 = -\partial\mathcal{L}/\partial y_1$, every iteration step satisfies the condition $y = y_1 + y_2$, thus learning on either y_1 or y_2 yields the same results: $(y_1 + \Delta y_1) + (y_2 + \Delta y_2) = (y_1 + \Delta y_1) + (y_2 - \Delta y_1) = y$.

3.2. Updating Scaling Factors

The contribution factors λ_i should be updated simultaneously with $\mathbf{Y}_i$. This can be accomplished by simply finding the maximum *a posteriori* values. Given the basis functions $\{\mathbf{W}_1, \mathbf{W}_2\}$ and the current estimate of the sources $\{\mathbf{Y}_1, \mathbf{Y}_2\}$, the posterior probability of λ_1 is

$$\begin{aligned} &P(\lambda_1|\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{W}_1, \mathbf{W}_2) \propto \\ &\quad P(\mathbf{Y}_1|\lambda_1, \mathbf{W}_1) P(\mathbf{Y}_2|1-\lambda_1, \mathbf{W}_2) p_\lambda(\lambda_1), \end{aligned} \qquad (14)$$

where $p_\lambda(\cdot)$ is the prior density function of λ. The value of λ_1 maximizing the posterior probability also maximizes the log of it,

$$\begin{aligned} \lambda_1^* &= \arg\max_\lambda \log P(\lambda|\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{W}_1, \mathbf{W}_2) \\ &= \arg\max_\lambda \{\mathcal{L} + \log p_\lambda(\lambda)\} , \end{aligned} \qquad (15)$$

where $\mathcal{L}$ is the log-likelihood of the estimated sources defined in equation 8. Assuming uniform distribution of λ in $[0, 1]$, $\partial\{\mathcal{L} + \log p_\lambda(\lambda)\}/\partial\lambda = \partial\mathcal{L}/\partial\lambda$, and it is calculated as

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = -\frac{\psi_1}{\lambda_1^2} + \frac{\psi_2}{\lambda_2^2} , \qquad (16)$$

where

$$\psi_i = \sum_{t=1}^{T_N} \varphi(\mathbf{s}_{it})^T \mathbf{W}_i \mathbf{y}_{it} \qquad (17)$$

derived by the chain rule

$$\frac{\partial \log p(\mathbf{s})}{\partial \lambda} = \frac{\partial \log p(\mathbf{s})}{\partial \mathbf{s}} \frac{\partial \mathbf{s}}{\partial \lambda} = \varphi(\mathbf{s})^T \mathbf{W} \mathbf{y} \left(-\frac{1}{\lambda^2} \right) . \qquad (18)$$

In the case of exponential power distributions, ψ is always less than or equal to zero because, for each coefficient s of $\mathbf{s}_{it}$ (subscripts are omitted for compact notation),

$$\begin{aligned}\varphi(s)\mathbf{w}_k\mathbf{y} &= \varphi(s)\frac{s}{\lambda} \\ &= -\frac{cq}{\sigma^q}|s|^{q-1}\mathrm{sign}(s)\frac{|s|\ \mathrm{sign}(s)}{\lambda} \\ &= -\frac{cq}{\sigma^q\lambda}|s|^q \ \leq\ 0 \qquad (19)\end{aligned}$$

$$\Leftrightarrow \psi = \sum_{t=1}^{T_N}\sum_{k=1}^{N}\varphi(s_{tk})\mathbf{w}_k\mathbf{y}_t \leq 0\,, \qquad (20)$$

where $\mathbf{w}_k$ is the k^{th} basis filter and equation 19 holds because $c, q, \sigma, \lambda \geq 0$. Therefore $\partial\mathcal{L}/\partial\lambda_1 = 0$ subject to $\lambda_1 + \lambda_2 = 1$ and $\lambda_1, \lambda_2 \in [0, 1]$ always has a solution at the local maxima of $\mathcal{L}$ such that

$$\frac{\partial\mathcal{L}}{\partial\lambda_1} = 0 \ \Leftrightarrow\ \frac{\lambda_1^2}{\lambda_2^2} = \frac{\psi_1}{\psi_2} \geq 0, \qquad (21)$$

$$\lambda_1^* = \frac{\sqrt{|\psi_1|}}{\sqrt{|\psi_1|}+\sqrt{|\psi_2|}}\,,\quad \lambda_2^* = \frac{\sqrt{|\psi_2|}}{\sqrt{|\psi_1|}+\sqrt{|\psi_2|}}\,. \qquad (22)$$

According to the above equation the algorithm updates the scaling factors w.r.t. the current estimate of the source signals.

4. EXPERIMENTAL RESULTS

We performed separation experiments on the simulated mixtures of four sound sources of different characteristics. They were monaural signals of rock and jazz music, male and female speech of the speakers 'mcpm0' and 'fdaw0' from the TIMIT speech database. Rock music was mainly composed of guitar and drum sounds, and jazz was generated by a wind instrument. Vocal parts of both music sounds were excluded. All signals were downsampled to 8kHz, from original 44.1kHz (music) and 16kHz (TIMIT speech). The following sections compare the characteristics of the sound sources by the learned basis functions in time domain and demonstrate the separation results. Audio files for all the experiments are accessible at the website **http://inc.ucsd.edu/~jangbal/ch1bss/**.

4.1. Comparing Basis Functions

Using the generalized Gaussian ICA learning algorithm we adapted the basis functions of the target sources. The training data are generated by employing every window of 64 samples long (8ms) starting at every sample of the source signals. The amount was approximately 7 seconds (56,000 datapoints) for each sound source. Figure 1 displays the examples of learned basis functions. Music basis functions exhibit consistent amplitudes with harmonics, and the speech basis functions changes their amplitudes along the time axis. Most of male basis functions are similar to a Gabor function (Gaussian modulated sinusoidal). See [10] for more detailed discussions about the speech basis. Figure 2 compares the four sound sources by average powerspectra of the basis functions. Each basis covers the frequency axis differently in amplitude, although there exists high degree of overlap especially between the two speech bases. These differences in time and frequency domain enable exclusion of each other which results in the recovered signals. We present the performed separation results in the following sections.

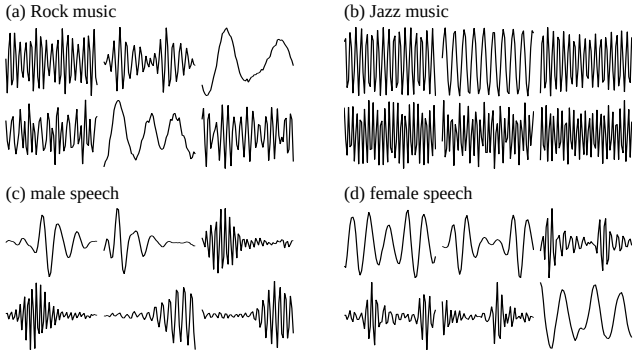

Fig. 1. Examples of learned basis functions for each sound sources. 6 was chosen and shown out of 64. The full set of basis functions is available on the website as well as the actual audio files. Basis functions of Jazz music are highly localized in frequency but not in time, and most of speech basis functions are localized both in time and frequency.

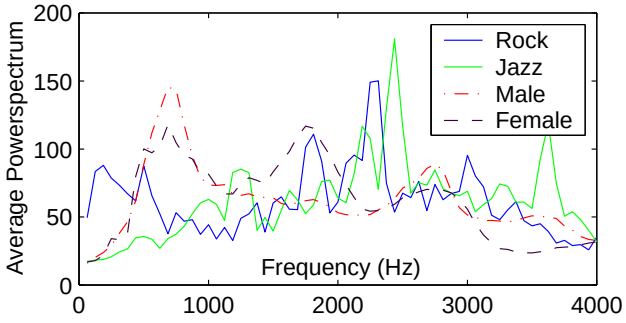

Fig. 2. Average powerspectra of the basis functions for each sound source. The powerspectra of the basis functions are averaged in each frequency band and represented in y-axis. Frequency scale ranges in 0∼4kHz (x-axis), since the data are sampled at 8kHz.

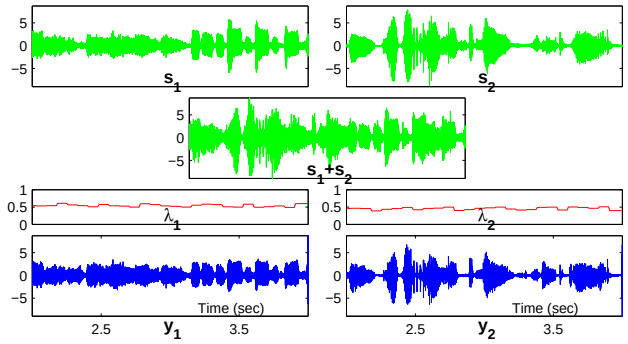

Fig. 3. Waveforms of separation results for the mixture of jazz music and male speech. In the vertical order the figures represent the time courses of: original sources ($\mathbf{S}_1$, $\mathbf{S}_2$), mixed signal ($\mathbf{S}_1+\mathbf{S}_2$), scaling factors blocked in 600 samples (λ_1, λ_2), and recovered signals ($\mathbf{Y}_1$, $\mathbf{Y}_2$). Left sides subscripted 1 are for the jazz music, and right sides subscripted 2 for the male speech.

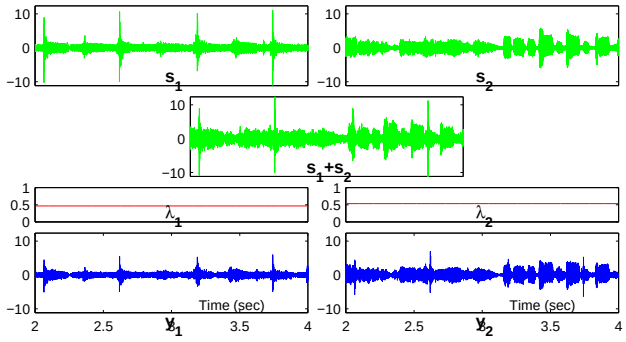

Fig. 4. Waveforms of separation results for the mixture of rock (left sides) and jazz music (right sides). Unlike jazz-male separation, scaling factors are set to be constant all over the time axis.

4.2. Separation Results

We generated single channel mixture by picking two sources out of the four and simply adding them. Then we applied the proposed method and reported the signal-to-noise ratio (SNR) of the recovered results (y_i: after separation) and that of the mixed (m: before separation) in table 1. Given the original source s and its estimate $\hat{s}$, SNR is defined by

$$\mathrm{snr}_s(\hat{s})\ [\mathrm{dB}] = 10\log_{10}\frac{\sum s^2}{\sum(s-\hat{s})^2}\,.$$

In terms of total SNR increase of both sources after separation, $\sum_{i=1}^{2}\{\mathrm{snr}_{s_i}(y_i) - \mathrm{snr}_{s_i}(m)\}$, sources are more cleanly recovered in the mixtures containing music signals than in the male-female mixture.

Table 1. SNR results. {R, J, M, F} stand for rock, jazz music, male, and female speech. *'Mixture'* columns are the sources that are mixed to m, and 'snr_{s_i}'s are the calculated SNR of mixed source (m) and recovered signals (y_i) given the original source (s_i).

Mixture	snr_{s_1}		snr_{s_2}		*Total*
$m = s_1 + s_2$	m	y_1	m	y_2	*increase*
R + J	-3.9	5.1	3.9	8.9	14.0
R + M	-3.7	3.5	3.7	7.2	10.7
R + F	-3.9	2.7	3.9	6.6	9.3
J + M	0.2	7.4	-0.2	7.2	**14.6**
J + F	0.0	6.5	0.0	6.6	13.1
M + F	-0.2	2.9	0.2	3.2	6.1

Separation of jazz music and male speech was the best, and the waveforms of source signals and separated signals are illustrated in figure 3. As shown in the third rows of the figure we updated the value of λ every block of 600 samples long, because amplitude changes frequently in the male speech. Note that λ is, though slightly, proportional to the amplitude of separated signal, which implies that blockwisely varying the value of λ enables temporal masking as in other CASA approaches.

Figure 4 demonstrates the separation results on the the rock-jazz mixture. Because music signals seldom change their amplitude, setting λ constant all over the time was better than blocking. For the other experimental results listed in table 1, audio files as well as waveform views are also available at the website.

5. EXPERIMENTS WITH REAL RECORDINGS

We have tested the performance of the proposed method on recordings in a real environment. The recorded signals are composed of a male speech utterance counting digits with music played in the background. The focus of this experiment is to recover the human speech in real recordings, to see how well the proposed separation algorithm works.

Data are recorded with two microphones, and we first applied the time-delayed blind deconvolution [12] and obtained the original sources. Then we used them as training data for the basis functions. The data are sampled at 8kHz, basis functions in 64 samples. The average powerspectra are compared in figure 5. Two kinds of sound sources have similar characteristics though low frequency components are more emphasized in background music sound. The algorithm

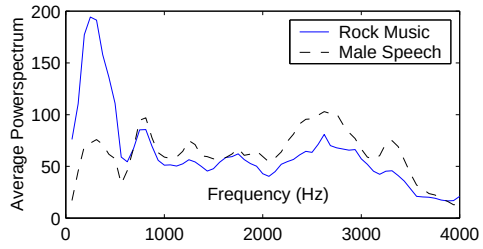

Fig. 5. Average powerspectra of the basis functions for music sound and male speech used in real data separation.

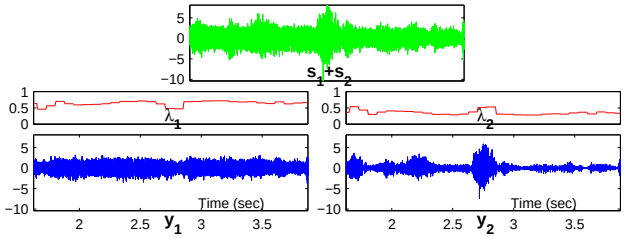

Fig. 6. Separation result for the real recording of mixed speech and music. Input signals are on top, below are λ_1 and λ_2 (blocked in 600 samples), and the recovered music and male speech.

successfully recovered the original sources as shown in figure 6. The full set of the basis functions as well as the separated results is available at the website also.

6. CONCLUSION

We presented a novel technique for single channel blind source separation. Instead of well-known prior knowledge of the sources, we exploited time-domain ICA basis functions that inherently capture the statistical structures of the sources. The algorithm recovers original auditory streams according to the gradient-ascent learning rule pursuing the maximum likelihood estimate of original sources, induced by the parameters of the basis filters and of the generalized Gaussian distributions of the filter coefficients. Separating two different sources was quite successful in the simulated mixtures of rock and jazz music, and male and female speech signals. Furthermore in the separation of the real recordings speech and background music were cleanly recovered. Consequently, the proposed method has many potential applications in real environments such as denosing for speech recognition and enhancement.

7. REFERENCES

[1] A. S. Bregman, *Computational Auditory Scene Analysis*, MIT Press, Cambridge MA, 1994.

[2] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1004–1034, 1995.

[4] H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Communications*, vol. 27, pp. 299–310, 1999.

[5] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. on Neural Networks*, vol. 10, pp. 684–697, 1999.

[6] S. T. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, vol. 13, pp. 793–799, 2001.

[7] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive-field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[8] A. J. Bell and T. J. Sejnowski, "Learning the higher-order structures of a natural sound," *Network: Computation in Neural Systems*, pp. 261–266, Jul 1996.

[9] M. S. Lewicki, "A flexible prior for independent component analysis," *submitted to Neural Computation*, 2000.

[10] T.-W. Lee and G.-J. Jang, "The statistical structures of male and female speech signals," in *Proc. ICASSP*, Salt Lake City, Utah, May 2001.

[11] B. Pearlmutter and L. Parra, "A context-sensitive generalization of ICA," in *Proc. ICONIP*, Hong Kong, Sept 1996, pp. 151–157.

[12] T.-W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," *Advances in Neural Information Processing Systems*, vol. 9, pp. 758–764, 1997.