THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.21, No.4E 2002. 12. pp. 146~155

Separation of Single Channel Mixture Using Time-domain Basis Functions

Gil-Jin Jang*, Yung-Hwan Oh*

*Department of Computer Science, Korea Advanced Institute of Science and Technology

(Received 8 April 2002; revised 18 June 2002; accepted 26 August 2002)

Abstract

We present a new technique for achieving source separation when given only a single channel recording. The main idea is based on exploiting the inherent time structure of sound sources by learning a priori sets of time-domain basis functions that encode the sources in a statistically efficient manner. We derive a learning algorithm using a maximum likelihood approach given the observed single channel data and sets of basis functions. For each time point we infer the source parameters and their contribution factors. This inference is possible due to the prior knowledge of the basis functions and the associated coefficient densities. A flexible model for density estimation allows accurate modeling of the observation, and our experimental results exhibit a high level of separation performance for simulated mixtures as well as real environment recordings employing mixtures of two different sources. We show separation results of two music signals as well as the separation of two voice signals.

Keywords: *Blind source separation, Independent component analysis (ICA), Computational auditory scene analysis (CASA), Adaptive filtering*

I. Introduction

The need for extracting individual sound sources from mixtures of different signals is increasing in both of the commercial and scientific fields. Many researchers in computational auditory scene analysis (CASA)[1] and independent component analysis (ICA)[2] formulated the problem as: we assume that the observed signal y^t is an addition of P independent source signals

$$y^t = \lambda_1 x_1^t + \lambda_2 x_2^t + \ldots + \lambda_P x_P^t, \quad (1)$$

where x_i^t is the t^{th} sampled value of the i^{th} source signal, and λ_i is the gain of each source which is fixed over time. In our manuscript, we denote superscripts for sample indices of time-varying signals and subscripts for the source identification. The gain constants are affected by several factors, such as powers, locations, directions and many other characteristics of the source generators as well as sensitivities of the sensors. It is convenient assumingall the sources to have zero mean and unit variance. The goal is to recover all x_i^t given only a single sensor input y^t. The problem is too ill conditioned to be mathematically tractable since the number of unknowns is $PT+P$ given only T observations.

Various sophisticated methods have been proposed in the research areas such as computational auditory scene

Corresponding author: Gil-Jin Jang (jangbal@speech.kaist.ac.kr)
Department of Computer Science, Korea Advanced Institute of Science and Technology, Gusong-dong, Usong-gu, Daejon 305-701, South Korea

analysis (CASA)[3,4] and independent component analysis (ICA)[2]. 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[3] or decorrelating frequency bands[4]. Recently Roweis[5] has presented a refiltering technique that 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.

ICA is a data driven method that relaxes the strong frequency characteristical assumptions. However, the ICA algorithms perform best when the number of the observed signals is greater than or equal to the number of sources [2]. Although some recent over-complete representations may relax this assumption, but the problem of separating sources from a single channel observation remains difficult. ICA has been shown to be highly effective in other aspects such as encoding speech signals[6] and natural sounds[7]. The basis functions and the coefficients learned by ICA constitute an efficient representation of the given time-ordered sequences of a sound source by estimating the maximum likelihood densities, thus reflecting the statistical structures of the sources.

This paper introduces a technique for separation of mixed sources in single channel observations 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. We make use of not only the ICA basis functions as a strong prior for the source characteristics, but their associated coefficient distributions modeled by generalized Gaussian density functions[8] as an objective function of the learning algorithm. 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.

II. Adapting Basis Functions and Model Parameters

The algorithm 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. We assume two different types of generative models in the observed single channel mixture as well as in the original sources. The first one is depicted in Figure 1-A. As described in Equation 1, at every $t \in [1, T]$ the observed instance is assumed to be a weighted sum of different sources. In our approach only the case of $P=2$ is regarded. This corresponds to the situation defined in Section 1 in that two different signals are mixed and observed in a single sensor.

For the individual source signals, we adopt a decomposition-based approach as another generative model. This approach was employed formerly in analyzing sound sources[6,7] by expressing a fixed-length segment drawn from a time-varying signal as a linear superposition of a number of elementary patterns, called basis functions, with scalar multiples (Figure 1-B). Contiguous samples of length N with $N<<T$ are chopped out of the whole samples of a source, from t to $t+N-1$, and the subsequent segment is denoted as an N-dimensional column vector in a boldface letter, $\mathbf{x}_i^t = [x_i^t \; x_i^{t+1} \; \cdots \; x_i^{t+N-1}]^T$, attaching the lead-off sample index for the superscript and representing the transpose operator with T. It is then expressed as a linear combination of the basis functions such that

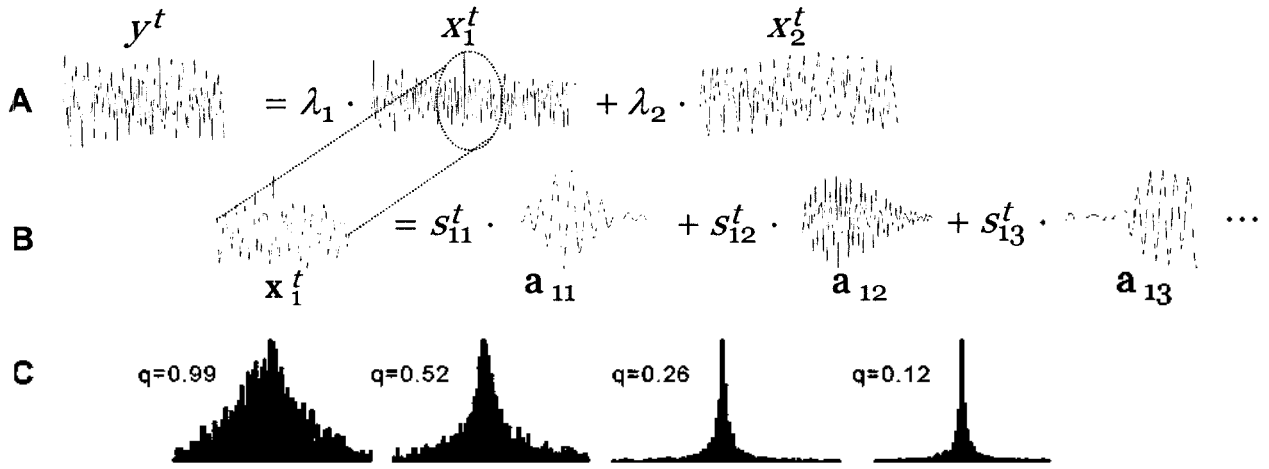

Figure 1. Generative models for the observed mixture and original source signals (A) A single channel observation is generated by a weighted sum of two source signals with different characteristics. (B) Individual source signals are generated by weighted linear superpositions of basis functions. (C) The distribution of the weight has a sharp summit at the mean, with long tails. They are modeled by generalized Gaussian density functions with varying the exponent.

$$\mathbf{x}_i^t = \sum_{k=1}^{M} \mathbf{a}_{ik} s_{ik}^t = \mathbf{A}_i \mathbf{s}_i^t \tag{2}$$

where M is the number of basis functions, $\mathbf{a}_{ik}$ is the k^{th} basis function of i^{th} source in the form of N-dimensional column vector, s_{ik}^t its coefficient (weight) and $\mathbf{S}_i^t = [s_{i1}^t\ s_{i2}^t \ldots s_{iM}^t]^T$. The r.h.s. is the matrix-vector notation. The second subscript k followed by the source index i in s_{ik}^t represents the component number of the coefficient vector $\mathbf{s}_i^t$. We assume that $M = N$ and $\mathbf{A}$ has full rank so that the transforms between $\mathbf{x}_i^t$ and $\mathbf{s}_i^t$ be reversible in both directions. The inverse of the basis matrix, $\mathbf{W}_i = \mathbf{A}_i^{-1}$, refers to the ICA filters that generates the coefficient vector, $\mathbf{s}_i^t = \mathbf{W}_i \mathbf{x}_i^t$.

The purpose of this decomposition is to model the multivariate distribution $P(\mathbf{x}_i^t)$ in a statistically efficient way. The ICA learning algorithm is equivalent to searching for the linear transformation that make the components as statistically independent as possible, as well as maximizing the marginal densities of the transformed coordinates for the given training data[9,10],

$$\mathbf{W}_i^* = \arg\max_{\mathbf{W}_i} \prod_t P(\mathbf{x}_i^t;\ \mathbf{W}_i) = \arg\max_{\mathbf{W}_i} \prod_t \prod_k P(s_{ik}^t). \tag{3}$$

Independence between the components factorizes the joint probabilities of the coefficients into the product of marginal ones, and independence over time does on the segments. What matters is therefore how well matched the model distribution is to the true underlying distribution of $P(s_{ik}^t)$. The coefficient histogram of real data reveals that the distribution has a highly sharpened point at the peak with a long tail (Figure 1-C). Therefore we use a generalized Gaussian prior[8] that provides an accurate estimate for symmetric non-Gaussian distributions by fitting the exponent q in the set of parameters θ in its simplest form

$$p(s \mid \theta) \propto \exp\left[-\left|\frac{s-\mu}{\sigma}\right|^q\right],\quad \theta = \{\mu, \sigma, q\}, \tag{4}$$

where $\mu = E[s]$ and $\sigma = \sqrt{V(s)}$. The ICA learning algorithm for the basis functions is proposed in[11], which incorporates generalized Gaussian priors in modeling source distributions. It can be derived using the maximum likelihood formulation. The probability density function of $\mathbf{x}_i^t$ is approximated by $\mathbf{W}_i$ and the density function of the coefficient vector, which is given by[12]:

$$P(\mathbf{x}_i^t) \cong p(\mathbf{s}_i^t \mid \Theta_i) \left|\det \mathbf{W}_i\right|, \tag{5}$$

where $p(\cdot)$ is the generalized Gaussian density function,

and $\Theta_i = \theta_{i,1\cdots M}$ — parameter group of all the coefficients, with the notation '$i\ldots j$' meaning an ordered set of the elements from index i to j. The term $|\det\ \mathbf{W}_i|$ gives the change in volume produced by the linear transformation $\mathbf{W}_i$ (see e.g.[13]). The log of Equation (5) is

$$\log P(\mathbf{x}_i^t) \cong \log p(\mathbf{s}_i^t \mid \Theta_i) + \log|\det\ \mathbf{W}_i|. \quad (6)$$

Maximizing the log-likelihood with respect to $\mathbf{W}_i$ gives a learning rule for $\mathbf{W}_i$[11]:

$$\Delta \mathbf{W}_i \propto [(\mathbf{W}_i^{\mathrm{T}})^{-1} - \varphi(\mathbf{s}_i^t)(\mathbf{x}_i^t)^{\mathrm{T}}], \quad (7)$$

where the component of the vector function $\varphi(s_i^t)$ is

$$\varphi(s) = \frac{\partial \log p(s \mid \theta)}{\partial s} = -\frac{cq}{\sigma^q}|s|^{q-1}\,\mathrm{sign}(s), \quad (8)$$

expressed by the parameters c, q, 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. An efficient way to maximize the log-likelihood is to follow the "natural" gradient[12],

$$\Delta \mathbf{W}_i \propto [(\mathbf{W}_i^{\mathrm{T}})^{-1} - \varphi(\mathbf{s}_i^t)(\mathbf{x}_i^t)^{\mathrm{T}}]\mathbf{W}_i^{\mathrm{T}}\ \mathbf{W}_i$$
$$= [\mathbf{I} - \varphi(\mathbf{s}_i^t)(\mathbf{s}_i^t)^{\mathrm{T}}]\ \mathbf{W}_i. \quad (9)$$

Here $\mathbf{W}_i^T\ \mathbf{W}_i$ rescales the gradient, simplifies the learning rule in Equation 7, and speeds convergence considerably. It has been shown that the general learning algorithm in Equation 9 can be derived from several theoretical viewpoints such as MLE[13], infomax[14], and negentropy maximization[2]. The basis filters $\mathbf{W}_i$ and their individual parameter set θ_{ik} are obtained beforehand and used as prior information for the following source separation algorithm.

III. Source Separation Algorithm

To infer the sources we perform log-likelihood maximization given the model parameters. Scaling factors of the generative model are learned as well.

3.1. Learning Rules for Source Signals

The learned basis filters maximize the likelihood of the given data. Suppose we were given the basis filters. Could we infer the learning data? The answer is generally "no" when $N < T$ and no other information is given. In our problem of single channel separation, half of the solution is already given by the constraint $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$, where x_i^t constitutes the basis learning data $\mathbf{x}_i^t$ (Figure 1-B). The goal of the algorithm to be presented is to complement the remaining half with the statistical information given by the coefficient density parameters θ_{ik} (Figure 1-C). At every time point a segment $\mathbf{x}_1^t$ generates the independent coefficient vector $\mathbf{s}_1^t = \mathbf{W}_1 \mathbf{x}_1^t$. Respectively $\mathbf{s}_2^t = \mathbf{W}_2 \mathbf{x}_2^t$. Assuming the independence over time, the probability of the whole signal $x_1^{1\cdots T}$ is determined by the marginal ones of all the possible segments, and can be computed by Equation (5)

$$P(x_1^{1\cdots T}) = \prod_{t=1}^{T_N} P(\mathbf{x}_1^t) \cong \prod_{t=1}^{T_N} P(\mathbf{s}_1^t \mid \Theta_1)|\det\ \mathbf{W}_1| \quad (10)$$

where, for convenience, $T_n = T - N + 1$. The objective function to be maximized is the multiplication of the total probability densities of both sound sources, and we denote its log by L:

$$L = \log P(x_1^{1\cdots T})P(x_2^{1\cdots T})$$
$$\cong \sum_{t=1}^{T_N}[\log p(\mathbf{s}_1^t \mid \Theta_1) + \log p(\mathbf{s}_2^t \mid \Theta_2)]$$
$$+ T_N \log|\det\ \mathbf{W}_1||\det\ \mathbf{W}_2|. \quad (11)$$

Our interest is in adapting x_1^t and x_2^t for $\forall t \in [1, T]$, toward the maximum of the objective function L. To infer the sound sources and their contribution factors simultaneously, instead of x_i^t we derive the learning rule on their weighted time-varying variables $z_i^t = \lambda_i x_i^t$, in a gradient-ascent manner by summing up the gradients of all the segments where the sample lies:

$$\frac{\partial l_z}{\partial z_1^t} = \sum_{n=1}^{N}\left[\frac{\partial}{\partial z_1^t}\log p(\mathbf{s}_1^{t_n} \mid \boldsymbol{\Theta}_1) + \frac{\partial}{\partial z_1^t}\log p(\mathbf{s}_2^{t_n} \mid \boldsymbol{\Theta}_2)\right]$$
$$= \sum_{n=1}^{N}\left[\sum_{k=1}^{N}\left\{\varphi(s_{1k}^{t_n})\frac{w_{1kn}}{\lambda_1}\right\} - \sum_{k=1}^{N}\left\{\varphi(s_{2k}^{t_n})\frac{w_{2kn}}{\lambda_2}\right\}\right]$$
$$\propto \sum_{n=1}^{N}\left[\lambda_2\sum_{k=1}^{N}\varphi(s_{1k}^{t_n})w_{1kn} - \lambda_1\sum_{k=1}^{N}\varphi(s_{2k}^{t_n})w_{2kn}\right], \quad (12)$$

which is derived by the fact that

$$\frac{\partial s_k^{t_n}}{\partial x^t} = \frac{\partial(\mathbf{w}_k\,\mathbf{x}^{t_n})}{\partial x^t}\frac{\partial x^t}{\partial z^t} = \frac{w_{kn}}{\lambda}$$

and

$$\frac{\partial z_2}{\partial z_1} = \frac{\partial(y - z_1)}{\partial z_1} = -1$$

where $t_n = t - n + 1$, $w_{ikn} = \mathbf{W}_i(k, n)$, and $\varphi(s)$ is defined in Equation 8. Note that the gradient of L for z_2, $\partial L/\partial z_2 = -\partial L/\partial z_1$, always makes the condition $y = z_1 + z_2$ satisfy, so learning rule on either z_1 or z_2 subsumes the other counterpart.

The detailed derivation of the proposed method is summarized as 4 stages in Figure 2. The whole figure shows one adaptation step of each sample. In stage (A), the source signals are decomposed into N statistically independent codes to be considered separately while obtaining the learning rules. The decomposition is done by a set of the given ICA filters. In (B), the stochastic gradient ascent for the filter output code is obtained by taking derivatives of the log probability density function of the code (Equation 8). In (C), the computed gradient is transformed to the source domain. All the filter output codes are regarded independent, so all the computations are performed independently. However in (D), we combine all the independently computed gradients and modify them to satisfy the initial constraint. The gradients are summed up and then scaled appropriately to obtain the final learning rates in Equation 12. The four stages comprise one iteration step. The solution is achieved after repeating this iteration on the source signal x_i^t to a convergence from a certain initial value.

Figure 3 gives a conceptual explanation for how to map Δs_{ik}^t to the original input domain Δx_i^t. Each $\mathbf{w}_{ik}$ takes windows of N contiguous samples, e.g. $\mathbf{x}_i^t = [x_i^t \ldots x_i^{t+N-1}]$, and using this as input the filter produces the source coefficient s_{ik}^t as output. Each individual sample

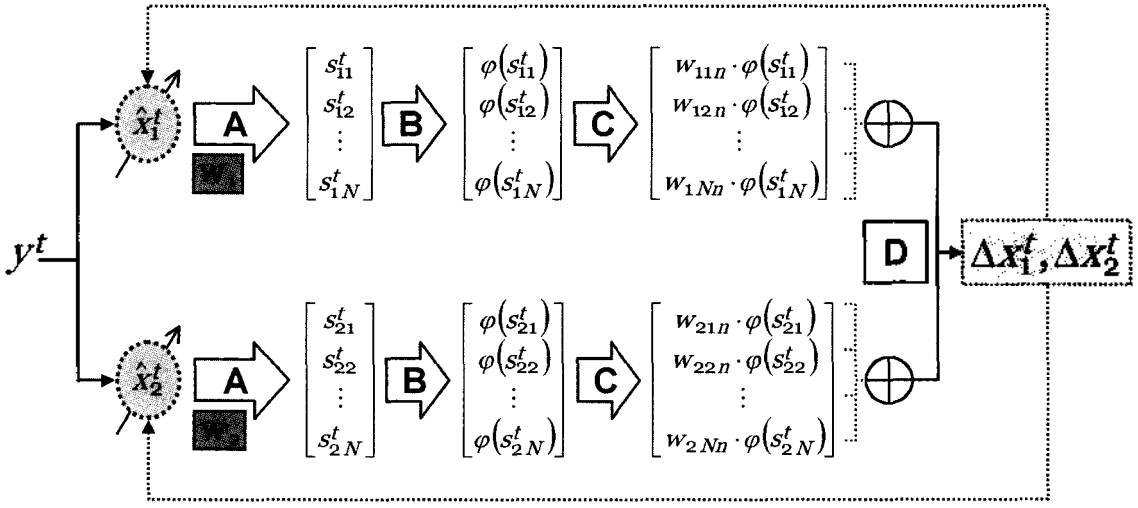

Figure 2. The overall structure and the data flow of the proposed method. In the beginning, we are given single channel data y^t, and we have the estimates of the source signals, $\hat{x}_i^t$, at every adaptation step. (A) $x_i^t \Rightarrow s_{ik}^t$: At each timepoint, the current estimates of the source signals are passed through a set of basis filters, generating N sparse codes s_{ik}^t that are statistically independent. (B) $s_{ik}^t \Rightarrow \Delta s_{ik}^t$: The stochastic gradient for each code is obtained by taking derivative of log likelihood of each individual code. (C) $\Delta s_{ik}^t \Rightarrow x_i^t$: The gradient for each code is transformed to the domain of source signal. (D) (finalization): The individual gradients are combined and modified to satisfy the given constraints, to be added to the current estimates of the source signals.

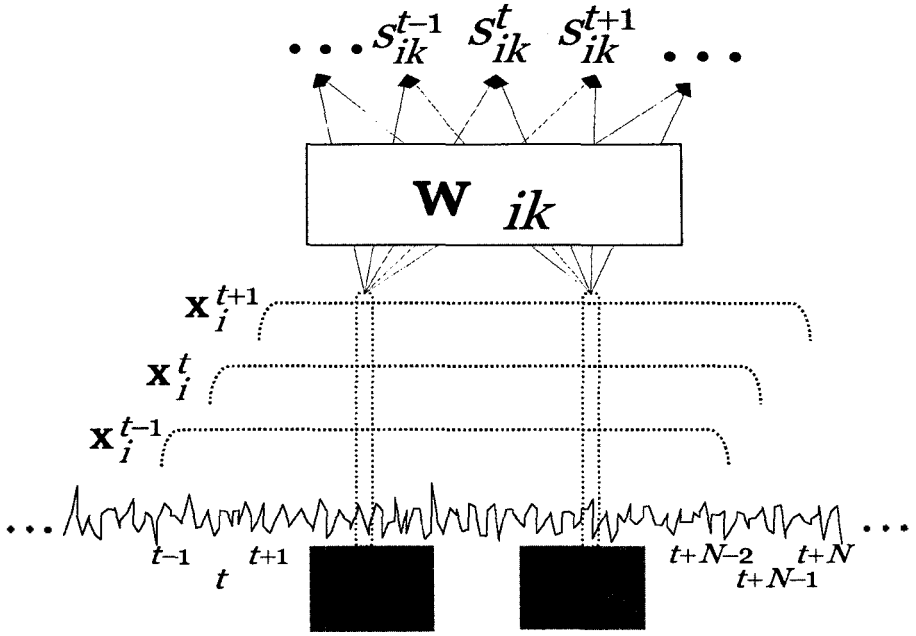

Figure 3. The participation of a sample in the source signal to the generation of each output coefficient. The input x_i^t is a vector composed of N contiguous samples ranging from t to $t+N-1$ in the sound source. The output coefficient s_{ik}^t is obtained by passing x_i^t through $\mathbf{w}_{ik}$, one of the basis filters. The middle of the figure shows that over a sample of the source signal there exist N different possible covers, which implies that a sample participates in the generation of N different output coefficients per filter.

of the source participates in the generation of N different inputs, and henceforth in the generation of N different output coefficients for each filter.

3.2. Updating Scaling Factors

Updating the contribution factors λ_i can be accomplished by simply finding the maximum a posteriori values. To simplify the inferring steps, we force the sum of the factors to be constant: e.g. $\lambda_1+\lambda_2=1$. Given the basis functions $\mathbf{W}_i$ and the current estimate of the sources $x_i^{1\cdots T}$, the posterior probability of λ_1 is

$$p(\lambda_1 \mid x_1^{1\cdots T}, x_2^{1\cdots T}; \mathbf{W}_1, \mathbf{W}_2) \propto P(x_1^{1\cdots T}; \mathbf{W}_1) P(x_2^{1\cdots T}; \mathbf{W}_2) p_\lambda(\lambda_1) \tag{13}$$

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

$$\lambda_1^* = \arg\max_\lambda \{L + \log p_\lambda(\lambda)\}, \tag{14}$$

where L is the log probability density of the estimated sources defined in Equation 11. Assuming that λ is uniformly distributed, $\partial\{L+\log p_\lambda(\lambda)\}/\partial\lambda = \partial L/\partial\lambda$, which is calculated as

$$\frac{\partial L}{\partial \lambda_1} = -\frac{\Psi_1}{\lambda_1^2} + \frac{\Psi_2}{\lambda_2^2}. \tag{15}$$

where

$$\Psi_i = \sum \frac{\partial \log p(\mathbf{s}_i^t \mid \Theta_i)}{\partial \lambda_i}, \tag{16}$$

derived by the chain rule

$$\frac{\partial \log p(\mathbf{s}_i^t)}{\partial \lambda_i} = \frac{\partial \log p(\mathbf{s}_i^t)}{\partial \mathbf{s}_i^t} \frac{\partial \mathbf{s}_i^t}{\partial \lambda_i} = \varphi(\mathbf{s}_i^t)^{\mathrm{T}} \mathbf{W}_i \mathbf{z}_i^t \left(-\frac{1}{\lambda_i^2}\right). \tag{17}$$

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

$$\varphi(s)\ \mathbf{w}_k\ z_i^t = \varphi(s)\frac{s}{\lambda}$$

$$= -q\frac{cq}{\sigma^q}|s|^{q-1}\,\mathrm{sign}(s)\frac{|s|\,\mathrm{sign}(s)}{\lambda}$$

$$= -q\frac{cq}{\sigma^q\lambda}|s|^q \le 0,$$

$$\Leftrightarrow \Psi = \sum_{t=1}^{T_N}\sum_{k=1}^{N}\varphi(s_k^t)\ \mathbf{w}_k\ z^t \le 0, \tag{18}$$

where $\mathbf{w}_k$ is the k^{th} basis filter and Equation 18 holds because $c,\ q,\ \sigma,\ \lambda \ge 0$. Therefore $\partial L/\partial\lambda_1 = 0$ subject to $\lambda_1 + \lambda_2 = 1$ and $\lambda_1, \lambda_2 \in [0,1]$ always has a solution at the local maxima of L such that

$$\frac{\partial L}{\partial\lambda_1} = 0 \Leftrightarrow \frac{\lambda_1^2}{\lambda_2^2} = \frac{\Psi_1}{\Psi_2} \ge 0, \tag{19}$$

$$\lambda_1^* = \frac{\sqrt{|\Psi_1|}}{\sqrt{|\Psi_1|}+\sqrt{|\Psi_2|}},\quad \lambda_2^* = \frac{\sqrt{|\Psi_2|}}{\sqrt{|\Psi_1|}+\sqrt{|\Psi_2|}}. \tag{20}$$

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

IV. Experimental Results

We have tested the performance of the proposed method on the single channel mixtures of four different sound types. They were monaural signals of rock and jazz music, male and female speech. We used different sets of speech signals for learning basis functions and for generating mixtures. For the mixture generation, two sentences of the target speakers 'mcpm0' and 'fdaw0', one for each, were selected from the TIMIT speech database. The training set consisted of 21 sentences for each gender, 3 for each of randomly chosen 7 males (or females) from the same database excluding the 2 target speakers. 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 8 kHz, from original 44.1 kHz (music) and 16 kHz (speech) data. The training data were segmented in 64 samples (8 ms) starting at every sample. Audio files for all the experiments are accessible at the website http://speech.kaist.ac.kr/~jangbal.

Figure 4 displays the actual source signals, some examples of adapted basis functions, and their coefficient densities. Music basis functions exhibit consistent amplitudes with harmonics, and the speech basis functions are similar to a Gabor wavelet, Gaussian modulated sinusoidal. Figure 5 compares four sound sources by the average powerspectra. Each covers all the frequency bands, although they are different in amplitude. One might expect that simple filtering or masking cannot separate the mixed sources clearly.

We generated single channel mixture by simply picking two sources out of the four and adding them. Then we applied the proposed method and reported the signal-to-noise ratios (SNRs) of the mixed signal (y^t: before separation) and the recovered results ($\hat{z}_i^t$: after separation) with the original sources ($z_i^t = \lambda_i x_i^t$) 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 the mixtures containing music signals are recovered more cleanly than male-female mixture. Separation of jazz music and male speech was the best, and the waveforms are illustrated in Figure 6.

In summary, our method has several advantages over traditional approaches to signal separation. They involve either spectral techniques[5] or time-domain nonlinear filtering techniques[3,4]. Spectral techniques assume that sources are disjoint in the spectrogram, which frequently result in audible distortions of the signal in the region where the assumption mismatches. Recent time-domain filtering techniques are based on splitting the whole signal space into several disjoint subspaces. Although they overcome the limit of spectral representation, they consider second-order statistics only, such as autocorrelation, which restricts the separable cases to orthogonal subspaces.

Our method avoids these strong assumptions by utilizing a prior set of basis functions that captures the inherent

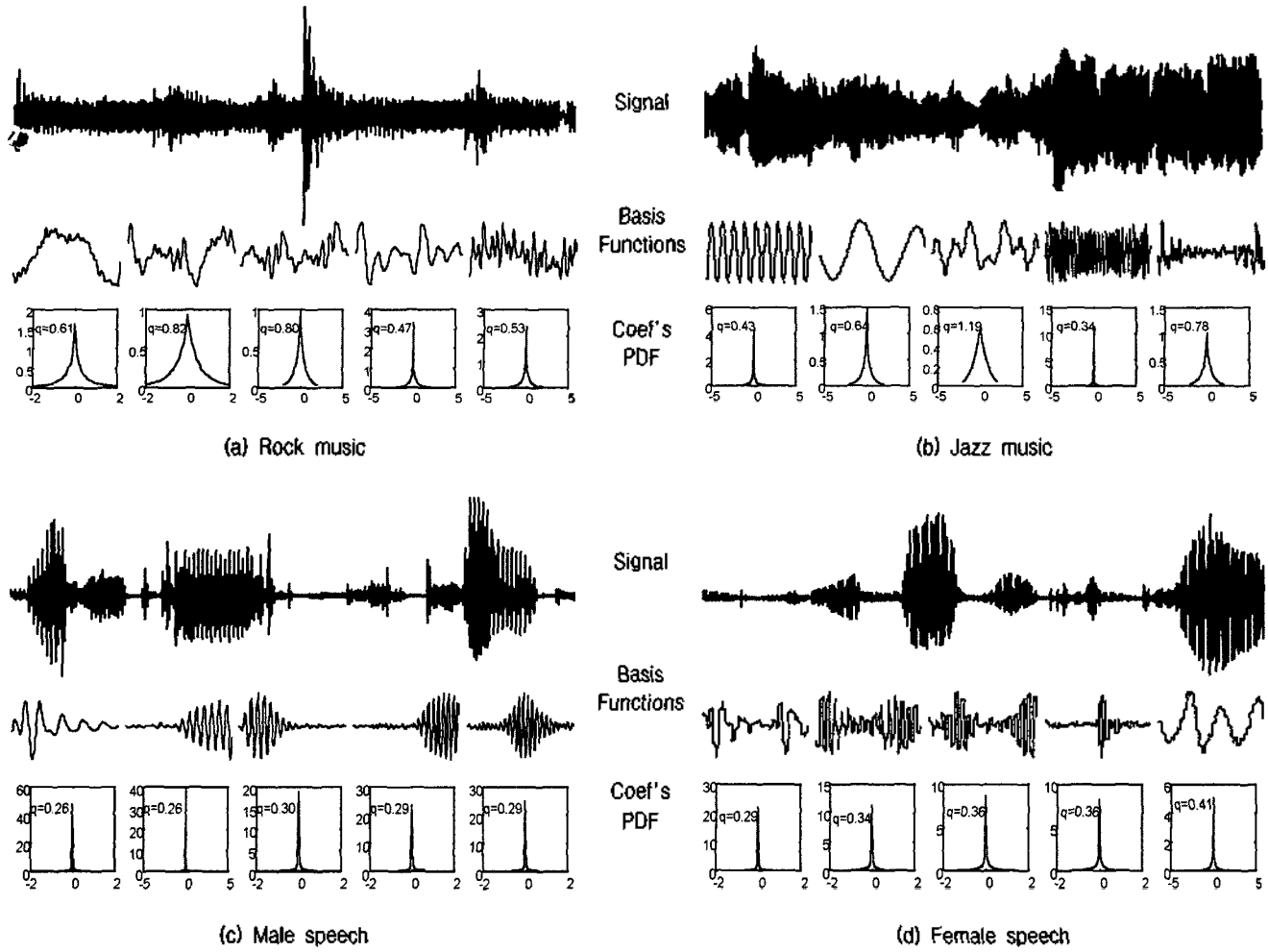

Figure 4. Waveforms of the 4 sound sources, learned basis functions (5 were chosen out of 64), and the estimated coefficient density functions. The full set of basis functions is available on the website as well.

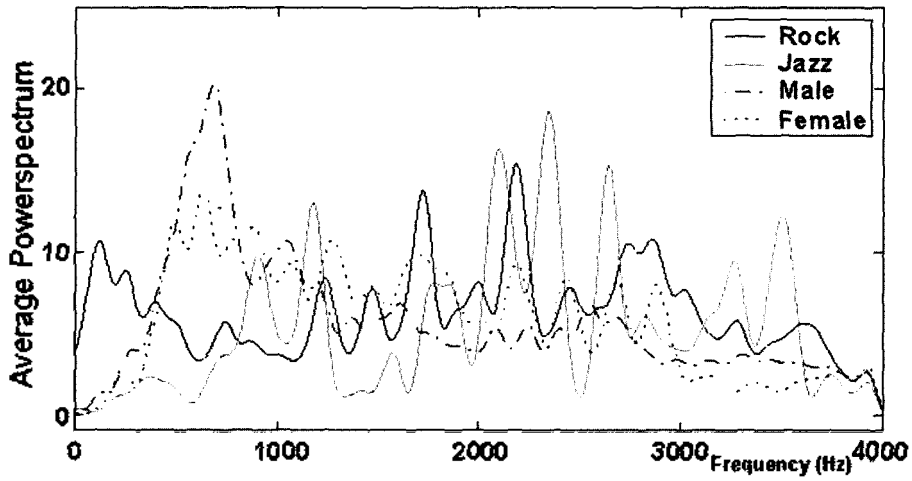

Figure 5. Average powerspectra of the 4 sound sources. Frequency scale ranges 0 to 4 kHz (x-axis), since all the signals are sampled at 8kHz. The powerspectra are averaged and represented in the y-axis.

statistical structures of the source signal. A decomposition-based, time-domain generative model on the source signals enables an iterative adaptation on the actual signals in the signal level. This generative model therefore makes use of spectral and temporal structures at the same time. The constraints are dictated by the ICA algorithm that forces the basis functions to result in an efficient representation, i.e. the linearly independent source coefficients; and both, the basis functions and their corresponding pdf are key to obtaining a faithful MAP based inference algorithm. An important question is how well the training data has to match the test data. We have also performed experiments with the set of basis functions learned from the test sounds and the SNR decreased on average by 1 dB.

V. Experiments with Real Recordings

We have tested the performance of the proposed method on the recordings in a real environment. Single microphone is used in recording a mixture of two sounds — mechanical noise and male speech. First we recorded the mere noise through the microphones without any intervening speech or other sound source, and used it as training data for the basis functions of the noise. Then we recorded a male speaker's talking under high level of noise, and applied our separation algorithm to obtain the two sources given only the single channel input. The basis functions of male speaker at the previous experiment (in Figure 4) are adopted for the recorded male speaker. The average powerspectra are compared in Figure 7. The two kinds of sound sources have similar characteristics though low frequency components of mechanical noise are more emphasized. The full set of the basis functions as well as the separated results is available at the website also. The algorithm successfully recovered the original sources as shown in Figure 8.

VI. Conclusions

We presented a technique for single channel source separation utilizing the time-domain ICA basis functions. Instead of well-known prior knowledge of the sources, we exploited the statistical structures of the sources that are inherently captured by the basis and its coefficients. The algorithm recovers original sound streams through gradient-ascent adaptation steps pursuing the maximum likelihood estimate of original sources, induced by the parameters of the basis filters and the generalized Gaussian distributions of the filter coefficients. With the separation results of the real recordings as well as simulated mixtures, we demonstrated that the proposed method is applicable to the real world problems such as source separation, denoising, and restoration of corrupted or lost data. Future research issues include providing a qualitative and objective description about the separability of our method.

References

1. G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, **8** (4), 297-336, 1994.
2. P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, **36**, 287-314, 1994.
3. H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Communications*, **27**, 299-310, 1999.
4. D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. on Neural Networks*, **10**, 684-697, 1999.
5. S. T. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, **13**, 793-799, 2001.
6. 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.
7. A. J. Bell and T. J. Sejnowski, "Learning the higher-order structures of a natural sound," *Network: Computation in Neural Systems*, 261-266, Jul. 1996.
8. S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis," *Journal of VLSI Signal Processing*, **26** (1-2), 25-38, 2000.
9. J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, **4**, 112-114, Apr. 1997.
10. B. Pearlmutter and L. Parra, "A context-sensitive generalization of ICA," *in Proc. ICONIP, (Hong Kong)*, 151-157, Sept. 1996.
11. T.-W. Lee and M. S. Lewicki, "The generalized gaussian mixture model using ICA," *in Proc. International Workshop on Independent Component Analysis (ICA'00), (Helsinki)*, 239-244, Jun. 2000.

12. S. Amari and J.-F. Cardoso, "Blind source separation — semiparametric statistical approach," *IEEE Trans. on Signal Proc.*, **45** (11), 2692-2700, 1997.
13. D. T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," *In Proc. EUSIPCO*, 771-774, 1992.
14. A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, **7**, 1129-1159, 1995.

[Profile]

• Gil-Jin Jang

Gil-Jin Jang received his B.S. and M.S. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, South Korea in 1997 and 1999 respectively. He is currently pursuing the Ph.D. degree in electrical engineering and computer science at KAIST. His research interests include statistical and adaptive signal processing, speech feature extraction, speech recognition, speech coding, and independent component analysis.

• Yung-Hwan Oh

Yung-Hwan Oh received his B.S. and M.S. degree from Seoul National University, South Korea and his Ph.D. degree from Tokyo Institute of Technology, Japan, 1972, 1974 and 1980 respectively. From 1981 to 1985, he was an assistant professor with the computer engineering department of Chungbuk National University. He was a visiting research staff at the University of California, Davis, from 1983 to 1984, and a visiting research professor at Carnegie-Mellon University from 1995 to 1996. He is now a professor with the department of electrical engineering and computer science of KAIST, Daejon, South Korea. His research interests are speech recognition, language identification, speech synthesis, speech coding, and speech enhancement.