Statistical Model of Speech Signals Based on Composite Autoregressive System with Application to Blind Source Separation

Hirokazu Kameoka, Takuya Yoshioka, Mariko Hamamura, Jonathan Le Roux, and Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation, 3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
{kameoka,leroux,kunio}@cs.brl.ntt.co.jp, takuya@cslab.kecl.ntt.co.jp

Abstract. This paper presents a new statistical model for speech signals, which consists of a time-invariant dictionary incorporating a set of the power spectral densities of excitation signals and a set of all-pole filters where the gain of each pair of excitation and filter elements is allowed to vary over time. We use this model to develop a combined blind separation and dereverberation method for speech. Reasonably good separations were obtained under a highly reverberant condition.

Keywords: Blind source separation, composite autoregressive system.

1 Introduction

When a set of observed data is considered to be a random sample drawn from a population that follows a particular family of probability distributions, we often refer to the distribution family as a statistical model. In this paper, we present a new statistical model of speech signals, suitable for applying to blind signal processing problems.

The aim of blind source separation (BSS) or blind dereverberation is to detect each source component from observed signals captured by one or several microphones without using any information about the transfer characteristic of the path from each source to each microphone. We thus usually make some assumption about the sources and then formulate an optimization problem based on a criterion that measures the consistency with the assumption. One approach to formulating the problem would be to define a likelihood function of a source signal by employing a statistical model assumption. When choosing which statistical model to invoke, it is important to take account of whether it agrees well with the actual behaviors of the source of interest and whether it leads to a mathematically tractable form of the optimization problem.

Let us now briefly review the BSS problem. BSS algorithms derived from a convolutive mixture model in the time domain such as [1] are fine for short mixing filters, but when it comes to realistically long filters, they can be unrealizable because of computational requirements. In the STFT domain where the frame

V. Vigneron et al. (Eds.): LVA/ICA 2010, LNCS 6365, pp. 245–253, 2010.

size is assumed to be sufficiently larger than the filter length, a convolutive mixture signal can be approximated by an instantaneous mixture in the frequency domain, thus allowing for an efficient implementation of BSS algorithms [2]. However, when reverberation comes into play, the filter length can often be larger than the frame size and so the above approximation becomes relatively less accurate. According to several studies such as [3], reverberation can be modeled fairly well as a convolution for each frequency-band in the STFT domain. Therefore, under highly reverberant conditions, we can expect a convolutive mixture model in the time-frequency domain [4] to be a better approximation.

We consider the situation where we observe signals emanating from M sources and captured by M microphones. Let $Y^l_{k,n}$ be the STFT of the signal observed at the l-th microphone and $\boldsymbol{Y}_{k,n} = (Y^1_{k,n}, \cdots, Y^M_{k,n})^{\mathrm{T}}$ be a set of observed data, where k and n are the frequency and time indices, respectively. We use $\boldsymbol{S}_{k,n} = (S^1_{k,n}, \cdots, S^M_{k,n})^{\mathrm{T}}$ to denote a set of M source components. In this paper, we use a separation system that has a multichannel finite-impulse-response form in the time-frequency domain such that:

$$\boldsymbol{S}_{k,n} = \sum_{\tau=0}^{n_k} \boldsymbol{W}_{k,\tau} \boldsymbol{Y}_{k,n-\tau}, \tag{1}$$

where $\boldsymbol{W}_{k,\tau},\ 0 \leq \tau \leq n_k$ are matrix coefficients of size $M \times M$. Let us assume that $S^m_{k,n}$ is a random variable and that $S^m_{k,n}$ and $S^{m'}_{k',n'}$ are statistically independent when $(k,n,m) \neq (k',n',m')$. The definition of the probability density function (pdf) for $S^m_{k,n}$, denoted by $f_{S^m_{k,n}|\theta^m}(s^m_{k,n}|\theta^m)$, is the main subject of the following section. This pdf corresponds to a statistical model of the m-th source signal and θ^m is the set of parameters characterizing its distribution. Once we define its specific form, the joint pdf of $\boldsymbol{Y} := \{\boldsymbol{Y}_{k,n}\}_{k,n}$ can be written in terms of $f_{S^m_{k,n}|\theta^m}(\ \cdot\ |\theta^m)$ explicitly as

$$f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}|\Theta) = \prod_k |\det \boldsymbol{W}_{k,0}| \prod_{t,m} f_{S^m_{k,n}|\theta^m}\left(\bar{S}^m_{k,n}|\theta^m\right), \tag{2}$$

where $\bar{S}^m_{k,n}$ is the m-th element of the separated signal vector, given by $\sum_\tau \boldsymbol{W}_{k,\tau} \boldsymbol{Y}_{k,n-\tau}$. The joint pdf $f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}|\Theta)$ is the likelihood of the unknown variables $\Theta := \{\{\theta^m\}_m, \{\boldsymbol{W}_{k,\tau}\}_{k,\tau}\}$ given observation $\boldsymbol{Y}$, which is an objective function for achieving separation and dereverberation in a joint manner, as with [4].

It has been shown that the speech production system can be well modeled on a frame-by-frame basis by a linear system comprising a glottal excitation input and a vocal-tract resonance filter that respectively determine the degree of periodicity and the phoneme of the voice. One of the most frequently used models for short-term speech signals is the autoregressive (AR) model, which models the signal as the output of an all-pole system. [5] was among the first to propose a BSS system in which a Gaussian AR source model is incorporated in $f_{S^m_{k,n}|\theta}(s^m_{k,n}|\theta^m)$, where a complex spectrum at each frame is assumed to be a set of random samples drawn from a different AR system with a Gaussian white noise input. This type of statistical model for STFT spectrograms of speech has

later been shown to work successfully for BSS in highly reverberant environments [4]. The objective of this paper is to investigate the possibility to improve the performance of this state-of-the-art BSS system [4] by replacing the Gaussian AR source model with the speech model we proposed previously [6], which will be reviewed in the next section.

2 Proposed Statistical Model of Speech

The white noise assumption as regards the excitation inputs underlying the standard AR model is known to not hold especially for voices with high fundamental frequencies (F_0s). This is because when F_0 increases the spacing between the harmonics of the excitation spectrum increases correspondingly, thus departing from a white (flat) spectrum. Rather than fixing the characteristics of the excitation input, we would therefore, if possible, like to estimate them in the same way as the vocal-tract characteristics. However, if we simply treat both of the characteristics as separate variables for each frame, it would no longer be possible to determine these characteristics uniquely since there are an infinite number of combinations giving the same filter output. Some additional constraint is needed to avoid this indeterminacy.

There are a wide variety of regularities in speech that can be exploited to constrain speech models. For example, it may be reasonable to assume that every short-term signal of a speech utterance can be represented by a combination of elements drawn from two dictionaries, one consisting of a small number of vocal-tract characteristics, and the other of a small number of excitation characteristics. This is because the phoneme number and periodicity range of speech during an entire utterance are both usually limited. We thus assume here that speech signals have been generated by a compound linear system composed of the direct product of a limited set of time-invariant excitation characteristics and a limited set of time-invariant vocal-tract characteristics where each output associated with an excitation and filter element pair is activated by a time-varying gain. A signal at a particular frame is thus assumed to be characterized by the volume levels of all the filter outputs. Note that since the dictionary elements are in general unknown, they need to be estimated in a data-driven manner.

In this section, we focus on a particular source and so the index m will be omitted for simplicity of notation. We start by modeling an output signal characterized by a particular excitation and filter element pair. Let us assume that a signal in the n-th frame, $\{x_n[t]\}_{t=1}^{K}$, is a sampled sequence drawn from the P-order AR process with a set of AR parameters common over n such that

$$x_n[t] = \sum_{p=1}^{P} a_p x_n[t-p] + \epsilon_n[t], \tag{3}$$

where $\epsilon_n[t]$ is an excitation input signal that is assumed to be a zero-mean stationary Gaussian noise. Its autocorrelation function, $\nu_n[t]$, is constant up to the gain over all the frames such that $\nu_n[t] = U_n h[t]$. U_n is assumed to be the

energy of $\epsilon_n[t]$ within the frame. We note here that $\epsilon_n[t]$ is not restricted to being white noise. Let $\boldsymbol{x}_n = (x_n[1], \cdots, x_n[K])^{\mathrm{T}} \in \mathbb{R}^K$ and its discrete Fourier transform (DFT) be $\boldsymbol{X}_n = (X_{1,n}, \cdots, X_{K,n})^{\mathrm{T}} \in \mathbb{C}^K$. Then, according to Eq. (3), $\boldsymbol{X}_n$ follows a multivariate complex Gaussian distribution $\boldsymbol{X}_n \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{\Lambda}_n)$, with a diagonal covariance matrix $\boldsymbol{\Lambda}_n = \mathrm{diag}\big(\lambda_{1,n}, \cdots, \lambda_{K,n}\big)$ whose elements are

$$\lambda_{k,n} = \frac{H_k U_n}{|A(e^{\mathrm{j}2\pi k/K})|^2}, \tag{4}$$

$$A(z) = 1 - a_1 z^{-1} - \cdots - a_P z^{-P}, \tag{5}$$

where k is the frequency index and j is the imaginary unit. $\{H_k\}_{k=1}^K$ is the DFT of $\{h[k]\}_{k=1}^K$, which represents the power spectral density (PSD) of the excitation source signal $\epsilon_n[t]$ (namely, the spectral fine structure), which can have any shape and is not necessarily flat. On the other hand, $1/|A(e^{\mathrm{j}2\pi k/K})|^2$ corresponds to a spectral envelope expressed as the PSD of the all-pole transfer function.

We can now construct a "composite" autoregressive model by extending the above model. The composite autoregressive system is assumed to consist of a dictionary of I excitation PSDs and a dictionary of J all-pole filters. Subsequently, we use superscripts i and j to denote the indices of the excitation PSDs and the all-pole filters, respectively, and we denote the ith excitation PSD and the jth all-pole transfer function by H_k^i and $1/A^j(e^{\mathrm{j}2\pi k/K})$. The system is able to generate $I \times J$ different voice components each of which is characterized by combining elements drawn from the respective dictionaries. If we assume that only one of the $I \times J$ voice components is active at each frame, the source pdf can be defined using a Gaussian scaled mixture model [7]. By contrast, we assume that all the voice components are always active with different volume levels. Let $U_n^{i,j}$ denote the volume level of the $\{i,j\}$-th voice component at the n-th frame. By following the discussion above, the DFT of the $\{i,j\}$-th voice component at the nth frame, $\boldsymbol{X}_n^{i,j} = (X_{1,n}^{i,j}, \cdots, X_{K,n}^{i,j})^{\mathrm{T}}$, follows a multivariate complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{\Lambda}_n^{i,j})$ with a diagonal covariance matrix $\boldsymbol{\Lambda}_n^{i,j} = \mathrm{diag}\big(\lambda_{1,n}^{i,j}, \cdots, \lambda_{K,n}^{i,j}\big)$ whose elements are

$$\lambda_{k,n}^{i,j} = \frac{H_k^i U_n^{i,j}}{|A^j(e^{\mathrm{j}2\pi k/K})|^2}, \tag{6}$$

$$A^j(z) = 1 - a_1^j z^{-1} - \cdots - a_P^j z^{-P}. \tag{7}$$

If we now assume that $\boldsymbol{X}_n^{1,1}, \cdots, \boldsymbol{X}_n^{I,J}$ are mutually independent, it follows that

$$\boldsymbol{S}_n = \sum_i \sum_j \boldsymbol{X}_n^{i,j} \sim \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{0}, \boldsymbol{\Phi}_n\big), \quad \boldsymbol{\Phi}_n = \sum_i \sum_j \boldsymbol{\Lambda}_n^{i,j}, \tag{8}$$

where $\boldsymbol{S}_n \in \mathbb{C}^K$ denotes the DFT of the speech signal at the n-th frame. The statistical model, $f_{S_{k,n}|\theta}(S_{k,n}|\theta)$, is thus given concisely as

$$f_{S_{k,n}|\theta}(S_{k,n}|\theta) = \frac{1}{\pi\phi_{k,n}} \exp\left(-\frac{|S_{k,n}|^2}{\phi_{k,n}}\right), \tag{9}$$

where θ contains all the unknown parameters of the present system: $\theta := \left\{H_k^i, a_p^j, U_n^{i,j}\right\}_{i,j,p,k,n}$. The diagonal element of $\boldsymbol{\Phi}_n$, i.e. $\phi_{k,n}$, corresponds to the PSD of the output signal produced by the present system such that

$$\phi_{k,n} = \sum_i \sum_j \frac{H_k^i U_n^{i,j}}{|A^j(e^{\mathrm{j}2\pi k/K})|^2}. \tag{10}$$

It is important to note that in the special case where $J = 1$ and $P = 0$, the PSD $\phi_{k,n}$ reduces to the form of matrix multiplication $\phi_{k,n} = \sum_i H_k^i U_n^{i,1}$ and Eq. (9) thus reduces to the likelihood function under the statistical interpretation of non-negative matrix factorization (NMF) given in [8]. This fact suggests that the present entire BSS system is structurally related to multichannel NMF [9].

3 Optimization Process

Given a set of observed STFT components $\boldsymbol{Y}$, we want to find the estimate of $\Theta = \{\{\theta^m\}_m, \{\boldsymbol{W}_{k,\tau}\}_{k,\tau}\}$ that maximizes the posterior density $f_{\Theta|\boldsymbol{Y}}(\Theta; \boldsymbol{Y}) \propto f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}; \Theta)\ f_\Theta(\Theta)$, or equivalently, the log posterior density

$$L(\theta) := \log f_{\boldsymbol{Y}|\Theta}(\boldsymbol{Y}|\Theta) + \log f_\Theta(\Theta). \tag{11}$$

Eq. (11) can be iteratively increased by using a coordinate descent method in which each iteration comprises the following three maximization steps: (S1) $\theta^m \leftarrow \mathrm{argmax}_{\theta^m} L(\theta)$ for all m, (S2) $\boldsymbol{W}_{k,0} \leftarrow \mathrm{argmax}_{\boldsymbol{W}_{k,0}} L(\theta)$ for all k, and (S3) $\{\boldsymbol{W}_{k,\tau}\}_{\tau=1}^{n_k} \leftarrow \mathrm{argmax}_{\{\boldsymbol{W}_{k,\tau}\}_{\tau=1}^{n_k}} L(\theta)$ for all k. If we were able to obtain an estimate of the PSD of each source, namely $\phi_{k,n}^m$, we could invoke [4] to perform (S2) and (S3). Therefore, obtaining the update formula of (S1) will suffice to complete the derivation of the entire optimization process. It should be noted that (S1) amounts to maximizing

$$\sum_{k,n} \log f_{S_{k,n}^m|\theta^m}(\bar{S}_{k,n}^m|\theta^m) + \log f_{\theta^m}(\theta^m) \tag{12}$$

with respect to θ_m where $\bar{S}_{k,n}^m$ is the m-th vector element of $\sum_\tau \boldsymbol{W}_{k,\tau}\boldsymbol{Y}_{k,n-\tau}$. As this maximization is carried out for each m separately, the index m is omitted again in the following.

$\log f_{S_{k,n}|\theta}(\bar{S}_{k,n}|\theta)$ is equal up to constant terms to the goodness of fit between $|\bar{S}_{k,n}|^2$ and $\phi_{k,n}$ defined by the Itakura-Saito divergence. We are thus led to obtain a PSD model with as small a reconstruction error as possible. On the other hand, as with the sparse coding concept [10], we would like to keep the voice components as sparse as possible. The prior term $\log f_\theta(\theta)$ can be used to promote the sparseness of $U_n^{i,j}$. In the subsequent analysis, for convenience we use an exponential prior (a folded Laplacian prior) defined over $U_n^{i,j} \geq 0$

$$f_\theta(\theta) = \prod_{i,j,n} \alpha \exp(-\alpha U_n^{i,j}), \tag{13}$$

which promotes sparsity when α is large. Maximizing Eq. (12) thus combines the goals of a small reconstruction error and sparseness. As a consequence, the more frequently a certain spectral fine/envelope structure emerges in $|\bar{S}_{k,n}|^2$, the more likely it is to be captured in the excitation/filter dictionary.

Although it is difficult to obtain a closed-form solution for maximizing Eq. (12), we can develop a computationally efficient scheme for its estimation based on the Generalized Expectation-Maximization (GEM) algorithm. When applying the GEM algorithm to the current "partial" MAP estimation problem (that is, (S1)), the first step is to define the "complete data". As the "observed" source component $\bar{S}_{k,n}$ is assumed to contain $I \times J$ concurrent voice components, a natural choice for the complete data is the corresponding hidden components, that is, $\boldsymbol{X}_{k,n} = (X_{k,n}^{1,1}, \cdots, X_{k,n}^{I,J})^{\mathrm{T}}$ with $X_{k,n}^{i,j} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{k,n}^{i,j})$. From Eq. (8), the many-to-one relationship between $\boldsymbol{X}_{k,n}$ and $\bar{S}_{k,n}$ is described as $\bar{S}_{k,n} = \mathbf{1}^{\mathrm{T}}\boldsymbol{X}_{k,n}$ with $\mathbf{1} = (1, \cdots, 1)^{\mathrm{T}}$. In Sec. 2, we have already assumed that $X_{k,n}^{i,j}$ is independent of all other $X_{k',n'}^{i',j'}$, so the log-likelihood of the complete data $\boldsymbol{X} := \{\boldsymbol{X}_{k,n}\}_{k,n}$ is

$$\log f_{\boldsymbol{X}|\theta}(\boldsymbol{X}|\theta) = -\sum_{k,n}\Big[\log\det \pi\boldsymbol{\Lambda}_{k,n} + \mathrm{tr}\big(\boldsymbol{\Lambda}_{k,n}^{-1}\boldsymbol{X}_{k,n}\boldsymbol{X}_{k,n}^{\mathrm{H}}\big)\Big], \tag{14}$$

where $\boldsymbol{\Lambda}_{k,n} = \mathrm{diag}\big(\lambda_{k,n}^{1,1}, \cdots, \lambda_{k,n}^{I,J}\big)$. Taking the conditional expectation of Eq. (14) given $\bar{S}_{k,n}$ and $\theta = \theta'$ and then adding $\log f_\theta(\theta)$ to both sides, we obtain

$$Q(\theta,\theta') = \log f_\theta(\theta) - \sum_{k,n}\Big[\log\det \pi\boldsymbol{\Lambda}_{k,n} + \mathrm{tr}\big(\boldsymbol{\Lambda}_{k,n}^{-1}\mathbb{E}\big[\boldsymbol{X}_{k,n}\boldsymbol{X}_{k,n}^{\mathrm{H}}|S_{k,n} = \bar{S}_{k,n}, \theta = \theta'\big]\big)\Big], \tag{15}$$

where $\mathbb{E}\big[\boldsymbol{X}_{k,n}\boldsymbol{X}_{k,n}^{\mathrm{H}}|S_{k,n} = \bar{S}_{k,n}, \theta = \theta'\big] = \boldsymbol{\Lambda}'_{k,n} - \boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big)^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n} + |\bar{S}_{k,n}|^2\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big)^{-1}\big(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}\mathbf{1}\big)^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}'_{k,n}$. Writing it in an element-wise expression, we obtain

$$Q(\theta,\theta') \stackrel{c}{=} -\sum_{k,n}\sum_{i,j}\left[\log H_k^i U_n^{i,j} + \frac{\Psi_{k,n}^{i,j}|A^j(e^{\mathrm{j}2\pi k/K})|^2}{H_k^i U_n^{i,j}}\right] - \alpha\sum_n\sum_{i,j}U_n^{i,j}, \tag{16}$$

where $\Psi_{k,n}^{i,j}$ represents the PSD estimate of the $\{i,j\}$th voice components,

$$\Psi_{k,n}^{i,j} = \frac{{\lambda'}_{k,n}^{i,j}}{{\phi'}_{k,n}}\big({\phi'}_{k,n} - {\lambda'}_{k,n}^{i,j} + \frac{{\lambda'}_{k,n}^{i,j}}{{\phi'}_{k,n}}|\bar{S}_{k,n}|^2\big). \tag{17}$$

The notation $\stackrel{c}{=}$ denotes equality up to constant terms. By setting the partial derivatives of $Q(\theta,\theta')$ with respect to H_k^i and $U_n^{i,j}$ at zero, we obtain the following update formulae for H_k^i and $U_n^{i,j}$

$$H_k^i = \frac{1}{NJ} \sum_n \sum_j \Psi_{k,n}^{i,j} |A^j(e^{\mathrm{j}2\pi k/K})|^2 / U_n^{i,j}, \tag{18}$$

$$\alpha U_n^{i,j\,2} + K U_n^{i,j} - \sum_k \Psi_{k,n}^{i,j} |A^j(e^{\mathrm{j}2\pi k/K})|^2 / H_k^i = 0, \quad U_n^{i,j} \geq 0 \tag{19}$$

By setting the partial derivatives of $Q(\theta, \theta')$ with respect to $a_1^j, \cdots, a_P^j$ at zero, we obtain the Yule-Walker equations

$$r_p^j = \sum_{q=1}^{P} a_q^j r_{p-q}^j \ (p = 1, \cdots, P), \tag{20}$$

where r_p^j is defined by the inverse DFT of the average spectral envelope over all the voice components with index j such that

$$r_p^j = \sum_k \sum_{n,i} \frac{\Psi_{k,n}^{i,j}}{H'^i_k U'^{i,j}_n} e^{\mathrm{p}j2\pi k/K}. \tag{21}$$

The update formula for the autoregressive parameters of the jth all-pole filter can therefore be calculated using the well-known Levinson-Durbin algorithm.

Here it is important to note that when a sparse constraint comes into play, there is a need for some constraint on the scales of the factorized elements in order to avoid an indeterminacy in the scaling. We thus adopt a simple procedure that consists of calculating Eq. (18) and then projecting it onto the unit norm space: $H_k^i \leftarrow H_k^i / \sum_{k'} H_{k'}^i$.

4 Experimental Results

We present here the separation results we obtained with the present BSS algorithm. All of the examples use the same two-input four-output impulse response, which was measured in a varechoic chamber where the reverberation time was 0.5 sec. With this impulse response, we mixed two speech signals into four mixtures. The two speech signals of female speakers, taken from the ATR speech database, were sampled at 16 kHz and band limited to the 50 Hz to 7 kHz frequency range. The input Signal-to-Interference ratios (SIRs) are shown in Tab. 1. Time-frequency representations were obtained using the polyphase filterbank analysis with a frame length of 32 ms and a hop size of 8 ms. The filter length n_k was set as follows: $n_k = 25$ for $F_k < 0.8$; $n_k = 20$ for $0.8 \leq F_k < 1.5$; $n_k = 15$ for $1.5 \leq F_k < 3$; $n_k = 10$ for $F_k \geq 3$, where F_k is the frequency in kHz of the k-th frequency bin. The AR order P was set at 12, and α at $10K$. The iterative algorithm comprising (S1)–(S3) was run for 3 iterations. For each step of (S1), the GEM algorithm was run for 100 iterations.

We tested the performance of the present method with different I and J settings. Tab. 2 lists the SIRs obtained with the proposed method for various settings of I and J, along with those obtained with the baseline methods, where

Table 1. Input SIRs (dB)

Source	Channel #1	#2	#3	#4
#1	-0.6	-0.3	-0.1	0.6
#2	0.6	0.3	0.1	-0.6

Table 2. Output SIRs (dB) for various settings

Source	$I=2$		$I=5$		$I=10$		Baseline1	Baseline2
	$J=12$	$J=15$	$J=8$	$J=10$	$J=8$	$J=12$		
#1	18.0	16.8	19.9	18.5	17.7	17.6	11.6	17.2
#2	11.7	10.9	14.1	13.9	13.6	14.5	11.0	13.9

Baseline1 and Baseline2 refer to Sawada's method [11] and Yoshioka's method [4], respectively. For Baseline1, we performed time-frequency analysis with a frame length of 256 ms and a hop size of 64 ms. For Baseline2, the frame length and the hop size were set at 16 ms and 8 ms. The best SIR result by the present method was obtained when I and J were set at 5 and 8, which significantly outperforms both of the baseline methods. The results are very preliminary and they need to be confirmed by a more thorough analysis in the future.

5 Conclusion

This paper described a statistical model called the "composite autoregressive system", which consists of a time-invariant dictionary incorporating a set of PSDs of excitation signals and a set of all-pole filters. Under this model, speech signals are assumed to be characterized by the volume levels of the excitation and filter element pairs, that vary over time. We proposed to use this model to develop a combined blind separation and dereverberation method for speech and reasonably good separations were obtained for four mixtures of two speech signals under a reverberant condition.

References

1. Douglas, S., Sawada, H., Makino, S.: Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters. IEEE Trans. Speech, Audio Process. 13(1), 92–104 (2005)
2. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neur. Comp. 22, 21–34 (1998)
3. Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.-H.: Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. In: Proc. Int'l. Conf. Acoust., Speech, Signal Process., pp. 85–88 (2008)
4. Yoshioka, T., Nakatani, T., Miyoshi, M., Okuno, H.G.: Blind separation and dereverberation of speech mixtures by joint optimization. IEEE Trans. Audio, Speech, Language Process (2010) (accepted for publication)
5. Dégerine, S., Zaïdi, A.: Separation of an instantaneous mixture of Gaussian autoregressive sources by the exact maximum likelihood approach. IEEE Trans. Signal Processing 52(6), 1499–1512 (2004)
6. Kameoka, H., Kashino, K.: Composite Autoregressive System for Sparse Source-Filter Representation of Speech. In: Proc. 2009 IEEE International Symposium on Circuits and Systems (ISCAS 2009), pp. 2477–2480 (2009)

7. Benaroya, L., Bimbot, F., Gribonval, R.: Audio source separation with a single sensor. IEEE Trans. Audio Speech Language Processing 14(1), 191–199 (2006)
8. Févotte, C., Bertin, N., Durrieu, J.-L.: Nonnegative matrix factorization,with the Itakura-Saito divergence. With application to music analysis. Neural Comput. 21(3), 793–830 (2009)
9. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. Audio, Speech, Language Process. 18(3), 550–563 (2010)
10. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
11. Sawada, H., Araki, S., Makino, S.: Measuring dependence of binwise separated signals for permutation alignment in frequency-domain BSS. In: Proc. Int'l. Symp. Circ., Syst., pp. 3247–3250 (2007)