Composite Autoregressive System for Sparse Source-Filter Representation of Speech

Hirokazu Kameoka, and Kunio Kashino
NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

***Abstract*— This paper presents a new generative model for speech signals called a "composite autoregressive system". This model consists of a composite dictionary incorporating a set of the power spectral densities (PSDs) of excitation sources 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 computationally efficient scheme for generating a sparse mixture representation of speech based on the Expectation-Maximization algorithm. The algorithm iteratively updates the excitation PSDs and the gains through the update formulae, which reduce under a particular condition to the multiplicative update rule for non-negative matrix factorization with the Itakura-Saito distance criterion, and the all-pole parameters using the Levinson-Durbin algorithm.**

I. Introduction

We live in a world containing a huge variety of sounds emanating from various combinations of sound sources. In recent years there has been a growing interest in the search for sparse representations of acoustic signals that allow us to analyze complex acoustic scenes in a data-driven manner. The assumption behind this approach is that the underlying model generating a time series comprising a wide variety of observations is simply a linear combination of a compact set of static dictionary vectors each of which is associated with a time-varying coefficient. Given a set of observed vectors (which can be either signals or spectra), the goal is to find a set of 'atoms', each defined by a dictionary vector, such that any observed vector can be parsimoniously represented using only a small subset of active atoms. As a consequence of this sparsity requirement, each of the atoms is capable of representing an independently recurrent pattern that underlies the observations. Sparse coding [1], non-negative matrix factorization (NMF) [2], shift-invariant sparse coding [3], shifted NMF [4], non-negative matrix factor deconvolution (NMFD) [5, 6], NMF2D [7], non-negative tensor factorization (NTF) [8], semi-NMF [9], shift-invariant semi-NMF [10], and complex NMF [11] can be designed to produce various types of sparse representations of acoustic signals, many of which have been applied extensively with notable success. The choice of which method to invoke depends on what kind of structural regularity or property we would like to exploit to obtain the desired form of signal decomposition.

This paper describes a signal model that allows us to obtain yet another type of sparse representation suited to analyzing speech. For this purpose we first consider factorizing speech into the product of prosodic and phonemic factors on the basis of the speech production mechanism. It has been shown that the speech production system can be well modeled by a linear system comprising a glottal excitation input and a vocal-tract resonance filter that respectively determine the degree of periodicity (*pitch*) and the phoneme (*timbre*) of the voice. Although the characteristic of a vocal-tract filter and the periodicity of a glottal excitation vary over time during an entire speech utterance, the phoneme number and periodicity range are both usually limited. It is therefore reasonable to assume that observed speech signals (of one or more concurrent voices) have been generated from a composite linear system that is composed of the direct product of a limited set of excitation sources and a limited set of vocal-tract filters where each excitation and filter element pair is associated with a time-varying activation coefficient that determines the instantaneous *loudness* of each of the concurrent voice components. We then assume that each vocal-tract filter is restricted to an all-pole filter. Since the output of an all-pole filter with a Gaussian white noise input is usually referred to as an autoregressive (AR) model, which is used in diverse applications of speech analysis, we will term the present generative model a "composite autoregressive system".

In the long term, we are planning to apply the present model to various blind signal processing problems including speech denoising, speech dereverberation and speech separation for which a time-varying all-pole filter model has proved to be very useful [14]. As a preliminary to these applications, this paper presents the basic formulation of the model, an efficient algorithm for estimating the model parameters, and some experimental results to show how well the present model represents speech signals.

II. Generative Model

A. Autoregressive system with colored noise input

We assume that a speech signal in a short-term segment, $\{x_t\}_{t=1}^K$, is a sampled sequence drawn from the P-order autoregressive process such that

$$x_t = \sum_{p=1}^{P} a_p x_{t-p} + \epsilon_t, \tag{1}$$

where ϵ_t is an excitation source signal that is assumed to be a zero-mean stationary Gaussian noise and its autocorrelation function is f_t. It is important to note that ϵ_t does not have to be a white noise. Let $\boldsymbol{x} = (x_1, \cdots, x_K)^{\mathrm{T}} \in \mathbb{R}^K$ and its discrete Fourier transform (DFT) be $\boldsymbol{X} = \boldsymbol{D}\boldsymbol{x} = (X_1, \cdots, X_K)^{\mathrm{T}} \in \mathbb{C}^K$ where $\boldsymbol{D}$ is a DFT matrix. Then, according to Eq. (1), the linearity of DFT, and the stationarity and the Gaussianity of ϵ_t, $\boldsymbol{X}$ follows a multivariate complex Gaussian distribution

$$\boldsymbol{X} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{\Lambda}), \tag{2}$$

978-1-4244-3828-0/09/$25.00 ©2009 IEEE

with a diagonal covariance matrix $\mathbf{\Lambda} = \mathrm{diag}\big(\lambda_1, \cdots, \lambda_K\big)$ whose diagonal elements are equal to

$$\lambda_k = \frac{F_k}{|A(e^{\mathrm{j}2\pi k/K})|^2}, \tag{3}$$

$$A(z) = 1 - a_1 z^{-1} - \cdots - a_P z^{-P}, \tag{4}$$

where k is the frequency index. F_k is the DFT of f_t and represents the power spectral density (PSD) of the excitation source signal ϵ_t (namely, the spectral fine structure), that can have any shape and is not necessarily flat. $1/|A(e^{\mathrm{j}2\pi k/K})|^2$, on the other hand, corresponds to the spectral envelope expressed as the PSD of the all-pole transfer function.

B. Composite autoregressive system

The composite autoregressive system is assumed to consist of a dictionary of I excitation source 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 let the ith excitation PSD and the jth all-pole transfer function be denoted by F_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. The gain of each voice component, i.e. the voice activity, is allowed to vary individually over time. We thus use $U_n^{i,j}$ to denote the voice activity of the $\{i,j\}$th voice component at the nth frame (or segment). Following the discussion in II-A, 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 $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Lambda}_n^{i,j})$ with a diagonal covariance matrix $\mathbf{\Lambda}_n^{i,j} = \mathrm{diag}\big(\lambda_{1,n}^{i,j}, \cdots, \lambda_{K,n}^{i,j}\big)$ whose elements are

$$\lambda_{k,n}^{i,j} = \frac{F_k^i U_n^{i,j}}{|A^j(e^{\mathrm{j}2\pi k/K})|^2}, \tag{5}$$

$$A^j(z) = 1 - a_1^j z^{-1} - \cdots - a_P^j z^{-P}. \tag{6}$$

Now, if we assume that $\boldsymbol{X}_n^{1,1}, \cdots, \boldsymbol{X}_n^{I,J}$ are mutually independent and i.i.d., it follows that

$$\boldsymbol{Y}_n = \sum_i \sum_j \boldsymbol{X}_n^{i,j} \sim \mathcal{N}_{\mathbb{C}}\big(\mathbf{0}, \mathbf{\Phi}_n\big), \ \mathbf{\Phi}_n = \sum_i \sum_j \mathbf{\Lambda}_n^{i,j}, \tag{7}$$

where $\boldsymbol{Y}_n \in \mathbb{C}^K$ denotes the DFT of the observed signal at the nth frame. Assuming the observations to be generated independently from the generative model, the probability density function (pdf) of $\boldsymbol{Y} = (Y_{k,n})_{K,N} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_N) \in \mathbb{C}^{K \times N}$ can be written concisely as $P(\boldsymbol{Y}|\theta) = \prod_{k,n} P(Y_{k,n}|\theta)$ with

$$P(Y_{k,n}|\theta) = \frac{1}{\pi \phi_{k,n}} \exp\left(-\frac{|Y_{k,n}|^2}{\phi_{k,n}}\right), \tag{8}$$

where θ contains all the unknown parameters of the system:

$$\theta = \bigcup_{i,j,k,n} \big\{F_k^i, a_p^j, U_n^{i,j}\big\}. \tag{9}$$

The diagonal element of $\mathbf{\Phi}_n$, i.e. $\phi_{k,n}$, corresponds to the PSD of the output signal produced by the system such that

$$\phi_{k,n} = \sum_i \sum_j \lambda_{k,n}^{i,j}. \tag{10}$$

III. Regularized Inverse Problem

A. Maximum A-Posteriori (MAP) estimation

Given a set of observed STFT components $\boldsymbol{Y}$, we would like to find the estimate of θ that maximizes the posterior density $P(\theta|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\theta)P(\theta)$, or equivalently,

$$\log P(\boldsymbol{Y}|\theta) + \log P(\theta). \tag{11}$$

As seen from an inspection of Eq. (8), $\log P(\boldsymbol{Y}|\theta)$ is equal up to constant terms to the goodness of fit between $|Y_{k,n}|^2$ and $\phi_{k,n}$ defined by the Itakura-Saito distance [12, 13]. 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 [1, 15], we would like to keep the voice activities as sparse as possible. For this purpose, the term $\log P(\theta)$ can be designed to have a sparsifying effect. In the subsequent analysis, for convenience we will use an inverse Gamma prior defined over $U_n^{i,j} > 0$

$$P(\theta) \propto \prod_{i,j,n} \frac{1}{U_n^{i,j\,\alpha+1}} \exp\left(-\frac{\beta}{U_n^{i,j}}\right), \tag{12}$$

which promotes sparsity when α is large. The reason for its convenience will become apparent later. Maximizing Eq. (11) therefore combines the goals of a small reconstruction error and sparseness. As a consequence, the more frequently a certain spectral fine/envelope structure emerges in the observed spectra, the more likely it is to be captured in the excitation/filter dictionary.

B. Expectation-Maximization (EM) algorithm

Although it is difficult to obtain a closed-form solution for the MAP estimate of θ, we can develop a computationally efficient scheme for its estimation based on the Expectation-Maximization (EM) algorithm [16]. In this subsection, we briefly review the general principle of the EM algorithm.

We use y to denote the observed "incomplete" data, that follows $P(y|\theta)$ indexed by the parameter θ, and let $\hat{y}$ denote the "complete" data, related to y by a certain noninvertible transform $H : \hat{y} \mapsto y$. The complete data pdf is written as

$$P(\hat{y}|\theta) = P(\hat{y}|y,\theta)P(y|\theta), \tag{13}$$

and, taking the logarithm on both sides, we obtain

$$\log P(y|\theta) = \log P(\hat{y}|\theta) - \log P(\hat{y}|y,\theta). \tag{14}$$

Now, taking the conditional expectation of both sides given y and the value $\theta = \theta'$, it follows that

$$\log P(y|\theta) = \mathbb{E}\big[\log P(\hat{y}|\theta)|y, \theta = \theta'\big] - \mathbb{E}\big[\log P(\hat{y}|y,\theta)|y; \theta = \theta'\big]. \tag{15}$$

Adding $\log P(\theta)$ to both sides and putting $L(\theta) \equiv \log P(y|\theta) + \log P(\theta)$, $Q(\theta,\theta') \equiv \mathbb{E}[\log P(\hat{y}|\theta)|y, \theta = \theta'] + \log P(\theta)$ and $V(\theta,\theta') \equiv \mathbb{E}[\log P(\hat{y}|y,\theta)|y, \theta = \theta']$, it reads

$$L(\theta) = Q(\theta,\theta') - V(\theta,\theta'). \tag{16}$$

We notice here that $L(\theta)$ has the same form as Eq. (11). By invoking the Jensen's inequality, $V(\theta,\theta') \leq V(\theta',\theta')$. Therefore, if we obtain θ such that $Q(\theta,\theta') > Q(\theta',\theta')$ then

$$L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}'). \tag{17}$$

Based on this result, we can construct an iterative algorithm that ensures the monotonic increase of $L(\theta)$ as follows:

Initial step: Set ℓ at 1 and $\theta^{(\ell)}$ at a certain value.
E step: Compute $Q(\theta, \theta^{(\ell)})$.
M step: Update θ by $\theta^{(\ell+1)} \leftarrow \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{(\ell)})$.
Set $\ell \leftarrow \ell + 1$ and return to *E step*.

C. Definition of incomplete data

When applying the EM algorithm to the current MAP estimation problem, the first step is to define the "complete data". As the observed STFT component $Y_{k,n}$ is assumed to contain $I \times J$ concurrent voice components, a natural choice for the complete data $\hat{\boldsymbol{Y}}_{k,n}$ is the corresponding hidden components, that is, $\hat{\boldsymbol{Y}}_{k,n} = (X_{k,n}^{1,1}, \cdots, X_{k,n}^{I,J})^{\mathrm{T}}$ with

$$X_{k,n}^{i,j} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{k,n}^{i,j}). \tag{18}$$

From Eq. (7), the many-to-one relationship between the complete data $\hat{\boldsymbol{Y}}_{k,n}$ and the incomplete data $Y_{k,n}$ is described as

$$Y_{k,n} = \boldsymbol{H}\hat{\boldsymbol{Y}}_{k,n}, \tag{19}$$

with $\boldsymbol{H} = \big[1, \cdots, 1\big] \in \mathbb{R}^{IJ}$. In II-B we have already seen that $X_{k,n}^{i,j}$ is mutually independent across i, j, k and n, so the log-likelihood of the complete data $\hat{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}}_{1,1}{}^{\mathrm{T}}, \cdots, \hat{\boldsymbol{Y}}_{K,N}{}^{\mathrm{T}})^{\mathrm{T}}$ is given by

$$\begin{aligned}\log P(\hat{\boldsymbol{Y}}|\theta) &= -\sum_{k,n}\Big[\log\det \pi\boldsymbol{\Lambda}_{k,n} + \hat{\boldsymbol{Y}}_{k,n}{}^{\mathrm{H}}\boldsymbol{\Lambda}_{k,n}^{-1}\hat{\boldsymbol{Y}}_{k,n}\Big] \\ &= -\sum_{k,n}\Big[\log\det \pi\boldsymbol{\Lambda}_{k,n} + \mathrm{tr}\big(\boldsymbol{\Lambda}_{k,n}^{-1}\hat{\boldsymbol{Y}}_{k,n}\hat{\boldsymbol{Y}}_{k,n}{}^{\mathrm{H}}\big)\Big], \end{aligned} \tag{20}$$

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. (20) given $Y_{k,n}$ and $\theta = \theta'$ and then adding $\log P(\theta)$ to both sides, we obtain

$$\begin{aligned}Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \log P(\theta) - \sum_{k,n}\Big[&\log\det \pi\boldsymbol{\Lambda}_{k,n} \\ &+ \mathrm{tr}\big(\boldsymbol{\Lambda}_{k,n}^{-1}\mathbb{E}\big[\hat{\boldsymbol{Y}}_{k,n}\hat{\boldsymbol{Y}}_{k,n}{}^{\mathrm{H}}|Y_{k,n}, \theta = \theta'\big]\big)\Big], \end{aligned} \tag{21}$$

where

$$\begin{aligned}&\mathbb{E}\big[\hat{\boldsymbol{Y}}_{k,n}\hat{\boldsymbol{Y}}_{k,n}{}^{\mathrm{H}}|Y_{k,n}, \theta = \theta'\big] = \\ &\boldsymbol{\Lambda}'_{k,n} - \boldsymbol{\Lambda}'_{k,n}\boldsymbol{H}^{\mathrm{T}}\big(\boldsymbol{H}\boldsymbol{\Lambda}'_{k,n}\boldsymbol{H}^{\mathrm{T}}\big)^{-1}\boldsymbol{H}\boldsymbol{\Lambda}'_{k,n} + \\ &|Y_{k,n}|^2\boldsymbol{\Lambda}'_{k,n}\boldsymbol{H}^{\mathrm{T}}\big(\boldsymbol{H}\boldsymbol{\Lambda}'_{k,n}\boldsymbol{H}^{\mathrm{T}}\big)^{-1}\big(\boldsymbol{H}\boldsymbol{\Lambda}'_{k,n}\boldsymbol{H}^{\mathrm{T}}\big)^{-1}\boldsymbol{H}\boldsymbol{\Lambda}'_{k,n}. \end{aligned} \tag{22}$$

Writing it in an elementwise expression and using c to denote the terms that do not depend on θ, we obtain

$$\begin{aligned}Q(\theta, \theta') = &-\sum_{k,n}\sum_{i,j}\left[\log F_k^i U_n^{i,j} + \frac{\Psi_{k,n}^{i,j}|A^j(e^{\mathrm{j}2\pi k/K})|^2}{F_k^i U_n^{i,j}}\right] \\ &-\sum_{n}\sum_{i,j}\left[(\alpha+1)\log U_n^{i,j} + \frac{\beta}{U_n^{i,j}}\right] + c, \end{aligned} \tag{23}$$

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

$$\Psi_{k,n}^{i,j} = {\lambda'}_{k,n}^{i,j}\left[1 + \frac{\lambda'^{i,j}_{k,n}\big(|Y_{k,n}|^2 - \phi'_{k,n}\big)}{\phi'^2_{k,n}}\right]. \tag{24}$$

D. M-step update formulae

We can derive closed-form expressions of the M-step update formulae for all the model parameters thanks to the definitions of the complete data and the prior. Differentiating $Q(\theta, \theta')$ partially with respect to F_k^i and $U_n^{i,j}$ and setting them at zero, we obtain the following update formulae

$$F_k^i = \frac{1}{NJ}\sum_n\sum_j \frac{\Psi_{k,n}^{i,j}|A^j(e^{\mathrm{j}2\pi k/K})|^2}{U_n^{i,j}}, \tag{25}$$

$$U_n^{i,j} = \frac{1}{K+\alpha+1}\left[\beta + \sum_k \frac{\Psi_{k,n}^{i,j}|A^j(e^{\mathrm{j}2\pi k/K})|^2}{F_k^i}\right]. \tag{26}$$

Differentiating $Q(\theta, \theta')$ partially with respect to $a_1^j, \cdots, a_P^j$ and setting them at zero, we obtain the normal equations

$$r_p^j = \sum_{q=1}^{P} a_q^j r_{p-q}^j \ (p = 1, \cdots, P), \tag{27}$$

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

$$r_p^j = \mathrm{DFT}^{-1}\left\{\sum_n\sum_i \frac{\Psi_{k,n}^{i,j}}{F_k^i U_n^{i,j}}\right\}. \tag{28}$$

The update formula for the autoregressive parameters of the jth all-pole filter can therefore be calculated via

$$\begin{bmatrix} a_1^j \\ \vdots \\ a_P^j \end{bmatrix} = \begin{bmatrix} r_0^j & \cdots & r_{1-P}^j \\ \vdots & \ddots & \vdots \\ r_{P-1}^j & \cdots & r_0^j \end{bmatrix}^{-1} \begin{bmatrix} r_1^j \\ \vdots \\ r_P^j \end{bmatrix}, \tag{29}$$

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. (25) and then projecting it onto the unit norm space: $F_k^i \leftarrow F_k^i / \sqrt{\sum_k F_k^{i2}}$ as with [15].

E. Reduction to NMF with Itakura-Saito distance

When $P = 0$, $J = 1$, and $P(\theta) = \mathrm{const.}$, the present algorithm reduces to the NMF algorithm with the Itakura-Saito distance criterion discussed thoroughly by Févotte in [13]. Since $P = 0 \Rightarrow |A^j(e^{\mathrm{j}2\pi k/K})|^2 = 1$, we verify that, under this condition, Eqs. (25), (26) become

$$F_k^i = \frac{1}{N}\sum_n \frac{\Psi_{k,n}^{i,1}}{U_n^{i,1}}, \quad U_n^{i,1} = \frac{1}{K}\sum_k \frac{\Psi_{k,n}^{i,1}}{F_k^i}, \tag{30}$$

which amount to the update formulae developed in [13].

IV. Experiment

In this section we report results for speech data excerpted from the ATR B-set speech database. Speech data was monaural and sampled at 16kHz. STFT was computed using a Hanning window that was 32ms long with a 16ms overlap. P was set at 12. The algorithm was run for 300 iterations. F_k^i, a_p^j and $U_n^{i,j}$ were initialized with random values.

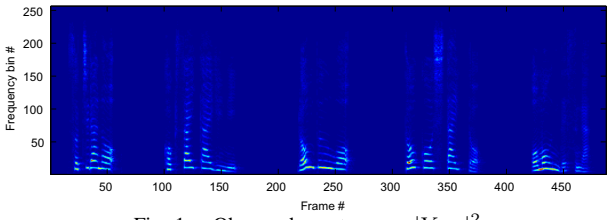

Fig. 1. Observed spectrogram $|Y_{k,n}|^2$.

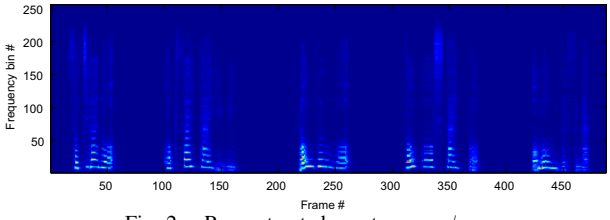

Fig. 2. Reconstructed spectrogram $\phi_{k,n}$.

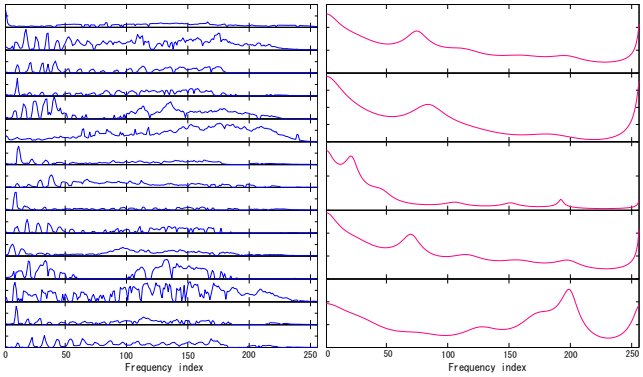

Fig. 3. Estimates of the excitation PSDs (left) and all-pole filters (right).

For the first experiment we tested the present method under various conditions for I and J on a single voice signal from a female speaker, whose spectrogram is shown in Fig. 1. For each condition we calculated the signal-to-noise ratio,

$$\mathrm{SNR(dB)} = 10\log_{10}\frac{\sum_{k,n}|Y_{k,n}|^2}{\sum_{k,n}\left||Y_{k,n}| - \sqrt{\phi_{k,n}}\right|^2}, \qquad (31)$$

to see how well the present model represented speech. The results can be seen in Tab. I. As examples, Fig. 2 and Fig. 3, respectively, show the reconstructed spectrogram and the dictionaries of the excitation source PSDs and all-pole filters obtained with the present method with $I = 15$ and $J = 5$.

For the second experiment we tested the present algorithm to ascertain whether the inverse Gamma prior had a sparsifying effect on the atomic activation. We used

$$\mathrm{sparseness}[n] = \frac{(\sum_{i,j}|U_n^{i,j}|^d)/(\sum_{i,j}U_n^{i,j\,2})^{d/2} - (IJ)^{1-d/2}}{1-(IJ)^{1-d/2}}$$

to measure the framewise sparseness of the voice activities $U_n^{i,j}$ across i and j. This function evaluates to unity if and only if $\{U_n^{i,j}\}_{i,j}$ contains only a single non-zero component, and takes a value of zero if and only if all components are equal. Fig. 4 shows the sparseness scores with $d = 1$ obtained with the present method and with the present method run without including the prior. The result confirmed that the inverse Gamma prior had a certain effect on the present model as regards obtaining a parsimonious mixture representation.

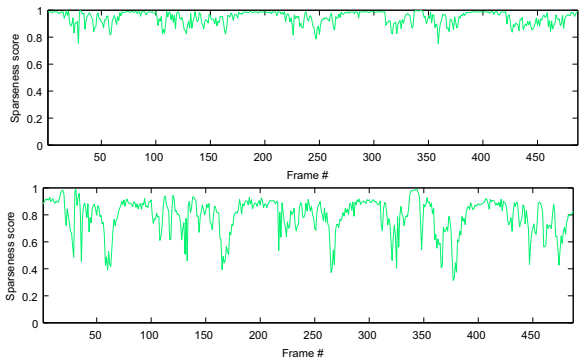

Fig. 4. Sparseness scores obtained with the present methods using an inverse Gamma prior (top) and using a uniform prior (bottom)

TABLE I

SIGNAL-TO-NOISE RATIO (DB) UNDER EACH CONDITION FOR I AND J.

$I \setminus J$	1	3	5
5	3.07	3.54	4.91
10	3.78	4.63	5.81
15	5.18	6.32	6.65

V. CONCLUDING REMARKS

We presented a new generative model for speech signals called a "composite autoregressive system". It consists of a composite dictionary of excitation source PSDs and all-pole filters where the gain of each excitation and filter element pair is allowed to vary over time. In addition, we used this model to develop a computationally efficient scheme for generating a sparse mixture representation based on the EM algorithm.

REFERENCES

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

[2] D.D. Lee, and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, **401**, pp. 788–791, 1999.

[3] M.S. Lewicki, and T.J. Sejnowski, "Coding time-varying signals using sparse, shift-invariant representations," Proc. NIPS'99, **11**, 1999.

[4] M. Mørup, K.H. Madsen, L.K. Hansen, "Shifted non-negative matrix factorization," Proc. MLSP'07, pp.139–144, 2007.

[5] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," Proc. SAPA'04, in CD-ROM, 2004.

[6] T. Virtanen, "Separation of sound sources by convolutive sparse coding," Proc. SAPA'04, in CD-ROM, 2004.

[7] M.N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," Proc. ICA'06, 2006.

[8] A. Shashua, T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," Proc. ICML'05, 792–799, 2005.

[9] C. Ding, T. Li, and M.I. Jordan, "Convex and semi-nonnegative matrix factorization for clustering and low-dimension representation," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-60428, 2006.

[10] J. Le Roux, A. de Cheveigné, and L.C. Parra, "Adaptive template matching with shift-invariant semi-NMF," Proc. NIPS'08, 2008.

[11] H. Kameoka, N. Ono, K. Kashino, S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," Proc. ICASSP'09, to appear.

[12] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," Proc. ICA'68, pp. C-17–C-20, 1968.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis," Tech. Rep. TELECOM ParisTech 2008D006, 2008.

[14] T. Yoshioka, T. Nakatani, and M. Miyoshi, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," Proc. EUSIPCO'08, in CD-ROM, 2008.

[15] P.O. Hoyer, "Non-negative sparse coding," Neural Networks for Signal Processing XII, pp. 557–565, 2002.

[16] M. Feder, E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," IEEE Trans. ASSP, **36**(4), pp. 477–489, 1988.