Single Microphone Blind Audio Source Separation Using EM-Kalman Filter and Short+Long Term AR Modeling

Siouar Bensaid, Antony Schutz, and Dirk T.M. Slock*

EURECOM
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
{siouar.bensaid,antony.schutz,dirk.slock}@eurecom.fr
http://www.eurecom.fr

Abstract. Blind Source Separation (BSS) arises in a variety of fields in speech processing such as speech enhancement, speakers diarization and identification. Generally, methods for BSS consider several observations of the same recording. Single microphone analysis is the worst underdetermined case, but, it is also the more realistic one. In this article, the autoregressive structure (short term prediction) and the periodic signature (long term prediction) of voiced speech signal are modeled and a linear state space model with unknown parameters is derived. The Expectation Maximization (EM) algorithm is used to estimate these unknown parameters and therefore help source separation.

Key words: blind audio source separation, EM, Kalman, speech processing, autoregressive.

1 Introduction

Blind Source Separation is an important issue in audio processing. It helps solving "'the cocktail party problem"' where each speaker needs to be retrieved independently. Several works exploit the temporal structure of speech signal to help separation. In literature, three categories can be listed : The first exploits only the short term correlation in speech signal and models it with a short term Auto-Regressive (AR) process [2]. A second category models the quasi-periodicity of speech by introducing the fundamental frequency (or pitch) in the analysis [3, 4]. Finally, few works combine the two aspects [5]. This article is classified in the last category. In [5], The problem is presented like an over-determined instantaneous model where the aim is to estimate jointly the long term (LT)and short term (ST) AR coefficients, as well as the demixing matrix in order to retrieve the speakers in a deflation scheme. An ascendant gradient algorithm is used to minimize the mean square of the total estimation error (short term and long term), and thus learn the parameters recursively. Our case is more difficult, since only a single sensor is used. Therefore, the proposed model

* EURECOM's research is partially supported by its industrial partners: BMW, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, ST Microelectronics, Swisscom, Thales. This research has also been partially supported by Project SELIA.

of speech propagation is rather simplified (the observation is the instantaneous sum of sources). Nevertheless, this model is still relevant in several scenarios. Using some mathematical manipulation, a state space model with unknown parameters is derived. Since the involved signals are Gaussians, Kalman filtering can be used in the EM algorithm (Expectation step) to estimate the state. This paper is organized as follows: The state space model is introduced in section 2. A small recapitulation of the EM-Kalman algorithm is presented in section 3 and the estimators' expressions are then computed. Numerical results are provided in section 4, and conclusions are drawn in section 5.

2 State Space Model Formulation

We consider the problem of estimating N_s mixed Gaussian sources. We use a voice production model [6], that can be described by filtering an excitation signal with long term prediction filter followed by a short term filter and which is mathematically formulated

$$\begin{aligned} y_t &= \sum_{k=1}^{N_s} s_{k,t} \; + \; n_t, \\ s_{k,t} &= \sum_{n=1}^{p_k} a_{k,n} \; s_{k,t-n} \; + \; \tilde{s}_{k,t} \\ \tilde{s}_{k,t} &= b_k \; \tilde{s}_{k,t-T_k} \; + \; e_{k,t} \end{aligned} \tag{1}$$

where

- y_t is the scalar observation.
- $s_{k,t}$ is the k^{th} source at time t, an AR process of order p_k
- $a_{k,n}$ is the n^{th} short term coefficient of the k^{th} source
- $\tilde{s}_{k,t}$ is the short term prediction error of the k^{th} source
- b_k is the long term prediction coefficient of the k^{th} source
- T_k is the period of the k^{th} source, not necessary an integer
- $\{e_{k,t}\}_{k=1..N_s}$ are the independent Gaussian distributed innovation sequences with variance ρ_k
- $\{n_t\}$ is a white Gaussian process with variance σ_n^2, independent of the innovations $\{e_{k,t}\}_{k=1..N_s}$

This model seems to describe more faithfully the speech signal, especially the voiced part (the most energetic part of speech). It is because it uses the short term auto-regressive model (AR) to describe the correlation between the signal samples jointly with the long term AR model to depict the harmonic structure of speech, rather than being restricted to just one of both [2, 3]. Let $\mathbf{x}_{k,t}$ be the vector of length $(N+p_k+2)$, defined like $\mathbf{x}_{k,t} = [s_k(t)\; s_k(t-1) \cdots s_k(t-p_k-1) \mid \tilde{s}_k(t)\; \tilde{s}_k(t-1) \cdots \tilde{s}_k(t-\lfloor T_k \rfloor) \cdots \tilde{s}_k(t-N+1)]^T$. This vector can be written in terms of $\mathbf{x}_{k,t-1}$ as the following

$$\mathbf{x}_{k,t} = \mathbf{F}_k \; \mathbf{x}_{k,t-1} \; + \; \mathbf{g}_k \; e_{k,t} \tag{2}$$

where $\mathbf{g}_k$ is the $(N+p_k+2)$ length vector defined as $\mathbf{g}_k = [\,1\,0\cdots0 \mid 1\,0\,\cdots\cdots\,0]^T$. The second non null component is at the position (p_k+3). The $(N+p_k+2)\times(N+p_k+2)$ matrix $\mathbf{F}_k$ has got the following structure

$$\mathbf{F}_k \;=\; \begin{bmatrix} \mathbf{F}_{11,k} & \mathbf{F}_{12,k} \\ \mathbf{O} & \mathbf{F}_{22,k} \end{bmatrix}$$

where the $(p_k+2)\times(p_k+2)$ matrix $\mathbf{F}_{11,k}$, the $(p_k+2)\times N$ matrix $\mathbf{F}_{12,k}$ and the $N\times N$ matrix $\mathbf{F}_{22,k}$ are given by

$$\mathbf{F}_{11,k} \;=\; \begin{bmatrix} a_{k,1} & a_{k,2} & \cdots & a_{k,p_k} & 0 & 0 \\ & & & & & \vdots \\ & & I_{(p_k+1)} & & & \vdots \\ & & & & & \vdots \\ & & & & & 0 \end{bmatrix}$$

$$\mathbf{F}_{12,k} \;=\; \begin{bmatrix} 0\cdots & (1-\alpha_k)\;b_k & \alpha_k\;b_k & 0\cdots & 0 \\ 0\cdots & 0 & 0 & 0\cdots & 0 \\ \vdots\;\ddots & \vdots & \vdots & \vdots\;\vdots & \vdots \\ 0\cdots & 0 & 0 & 0\cdots & 0 \end{bmatrix}$$

$$\mathbf{F}_{22,k} \;=\; \begin{bmatrix} 0\cdots & (1-\alpha_k)\;b_k & \alpha_k\;b_k & 0\cdots & 0 \\ & & & & \vdots \\ & & I_{(N-1)} & & \vdots \\ & & & & \vdots \\ & & & & 0 \end{bmatrix}$$

In matrices $\mathbf{F}_{12,k}$ and $\mathbf{F}_{22,k}$, the variable α_k, given by $\alpha_k \;=\; (1-\frac{\lfloor(T_k)\rfloor}{T_k})$, is present to consider the case where pitches are not integer. It is noteworthy that the choice of the $\mathbf{F}_{22,k}$ matrix size N should be done carefully. In fact, the value of N should be superior to the maximum value of pitches T_k in order to detect the long-term aspect. It can be noticed that the coefficients $(1-\alpha_k)\;b_k$ and $\alpha_k\;b_k$ are situated respectively in the $\lfloor T_k\rfloor^{th}$ and $\lceil T_k\rceil^{th}$ columns of $\mathbf{F}_{22,k}$ and $\mathbf{F}_{12,k}$. Since N_s sources are present, we introduce the vector $\mathbf{x}_t$ that consists of the concatenation of the $\{\mathbf{x}_{k,t}\}_{k=1:N_s}$ vectors ($\mathbf{x}_t \;=\; [\mathbf{x}_{1,t}^T\;\mathbf{x}_{2,t}^T\;\cdots\;\mathbf{x}_{N_s,t}^T]^T$) which results in the time update equation 3. Moreover, by reformulating the expression of $\{y_t\}$, we introduce the observation equation 4. We obtain the following state space model

$$\mathbf{x}_t = \mathbf{F}\;\mathbf{x}_{t-1} \;+\; \mathbf{G}\;\mathbf{e}_t \tag{3}$$

$$y_t = \mathbf{h}^T\mathbf{x}_t \;+\; n_t \tag{4}$$

where

- $\mathbf{e}_t = [e_{1,t}\;e_{2,t}\;\cdots\;e_{N_s,t}]^T$ is the $N_s\times1$ column vector resulting of the concatenation of the N_s innovations. Its covariance matrix is the $N_s\times N_s$ diagonal matrix $\mathbf{Q}$ =diag$(\rho_1,\cdots,\rho_{N_s})$.

- $\mathbf{F}$ is the $\sum_{k=1}^{N_s}(p_k+N+2) \times \sum_{k=1}^{N_s}(p_k+N+2)$ block diagonal matrix given by $\mathbf{F} = \text{blockdiag}\,(\mathbf{F}_1, \cdots, \mathbf{F}_{N_s})$.
- $\mathbf{G}$ is the $\sum_{k=1}^{N_s}(p_k+N+2) \times N_s$ matrix given by $\mathbf{G} = \text{block diag}(\mathbf{g}_1, \cdots, \mathbf{g}_{N_s})$
- $\mathbf{h}$ is the $\sum_{k=1}^{N_s}(p_k+N+2) \times 1$ column vector given by $\mathbf{h} = [\mathbf{h}_1^T \cdots \mathbf{h}_{N_s}^T]^T$ where $\mathbf{h}_i = [1\ 0 \cdots\ 0]^T$ of length $(N+p_k+2)$.

It is obvious that the linear dynamic system derived before depends on unknown parameters recapitulated in the variable $\theta = \Big\{ \{a_{k,n}\}_{\substack{k\in\{1,\ldots,N_s\} \\ n\in\{1,\ldots,p_k\}}}, \{b_k\}_{k\in\{1,\ldots,N_s\}}, \{\rho_k\}_{k\in\{1,\ldots,N_s\}}, \sigma_n^2 \Big\}$. Hence, a joint estimation of sources (the state) and θ is required. In literature([11, 10, 7]), the EM-Kalman algorithm presents an efficient approach for estimating iteratively parameters and its convergence to the Maximum Likelihood solution is proved [9]. In the next section, the application of this algorithm to our case is developed.

3 EM-Kalman Filter

The EM-Kalman algorithm permits to estimate iteratively parameters and sources by alternating two steps : E-step and M-step [9]. In the M-step, an estimate of the parameters $\hat{\theta}$ is computed. In our problem, there are two types of parameters: the parameters of the time update equation 3 which consist on the short term and long term coefficients and the innovation power of all the N_s sources, and one parameter of the observation equation 4, the observation noise power. From the state space model presented in the first part, and for each source k, the relation between the innovation process at time $t-1$ and the LT+ST coefficients could be written as

$$e_{k,t-1} = \mathbf{v}_k^T \breve{\mathbf{x}}_{k,t-1} \tag{5}$$

where $\mathbf{v}_k = [1\ -a_{k,1} \cdots -a_{k,p_k}\ -(1-\alpha_k)\ b_k\ -\alpha_k\ b_k]^T$ is a $(p_k+3)\times 1$ column vector and $\breve{\mathbf{x}}_{k,t-1} = [s_k(t-1,\theta) \cdots s_k(t-p_k-1,\theta)\ \tilde{s}_k(t-\lfloor T_k \rfloor -1,\theta)\ \tilde{s}_k(t-\lfloor T_k \rfloor -2,\theta)]^T$ is called the partial state deduced from the full state $\mathbf{x}_t$ with the help of a selection matrix $\mathbf{S}_k$. This lag of one time sample between the full and partial state is justified later. After multiplying (5) by $\breve{\mathbf{x}}_{k,t-1}^T$ in the two sides, applying the operator $E\{\ |y_{1:t}\}$ and doing a matrix inversion, the following relation between the vector of coefficients and the innovation power is deduced

$$\mathbf{v}_k = \rho_k \mathbf{R}_{k,t-1}^{-1} [1, 0 \cdots 0]^T \tag{6}$$

where the covariance matrix $\mathbf{R}_{k,t-1}$ is defined as $E\Big\{\breve{\mathbf{x}}_{k,t-1}\breve{\mathbf{x}}_{k,t-1}^T | y_{1:t}\Big\}$. It is important to notice that the estimation of $\mathbf{R}_{k,t-1}$ is done using observations till time t, which consists on a fixed-lag smoothing treatment with $lag = 1$. As mentioned previously, the relation between the partial state at time $t-1$ and the full state at time t is $\breve{\mathbf{x}}_{k,t-1} = \mathbf{S}_k \mathbf{x}_t$. This key relation is used in the partial state covariance matrix computation

$$\mathbf{R}_{k,t-1}^{-1} = \mathbf{S}_k E\Big\{\mathbf{x}_t \mathbf{x}_t^T | y_{1:t}\Big\} \mathbf{S}_k^T \tag{7}$$

Notice here the transition from the fixed lag smoothing with the partial state to the simple filtering with the full state. This fact justifies the selection of the partial state at time $t-1$ from the full state at time t. This selection is possible due to the augmented form matrix F_k or more precisely $\mathbf{F}_{11,k}$. The innovation power is simply deduced as the first component of the matrix $\mathbf{R}_{k,t-1}^{-1}$. The estimation of the observation noise power σ_n^2 is achieved by maximizing the loglikewood function $\log P\left(y_t|\mathbf{x}_t,\sigma_n^2\right)$ relative to σ_n^2. The optimal value can be easily proved equal to

$$\hat{\sigma_n^2}^{(t)} = y_t^2 - 2y_t\mathbf{h}^T\hat{\mathbf{x}}_{t|t} + \mathbf{h}^T\Big(\hat{\mathbf{x}}_{t|t}\hat{\mathbf{x}}_{t|t}^T + \mathbf{P}_{t|t}\Big)\mathbf{h} \tag{8}$$

The time index in (t) in $\hat{\sigma_n^2}^{(t)}$ is to denote the iteration number. The computation of the partial covariance matrix $\mathbf{R}_{k,t-1}$ is achieved in the $E-step$. This matrix depends on the quantity $E\Big\{\mathbf{x}_{k,t}\mathbf{x}_{k,t}^T|y_{1:t}\Big\}$ the definition of which is

$$E\Big\{\mathbf{x}_t\mathbf{x}_t^T|y_{1:t}\Big\} = \hat{\mathbf{x}}_{t|t}\hat{\mathbf{x}}_{t|t}^T + \mathbf{P}_{t|t} \tag{9}$$

where the quantities $\hat{\mathbf{x}}_{t|t}$ and $\hat{\mathbf{P}}_{t|t}$ are respectively the full estimated state and the full estimation error covariance computed using Kalman filtering equations. The adaptive algorithm is presented as Algorithm 1. The algorithm needs an accurate initialization, which will be discussed afterward. In the algorithm $\hat{\mathbf{s}}_{k,t}$ is the estimation of the source k at time t.

Adaptive EM Kalman Algorithm

– E-Step. Estimation of the sources covariance

$$\begin{aligned}
\mathbf{K}_t &= \mathbf{P}_{t|t-1}\mathbf{h}(\mathbf{h}^T\mathbf{P}_{t|t-1}\mathbf{h} + \hat{\sigma_n^2})^{-1}\\
\hat{\mathbf{x}}_{t|t} &= \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{h}^T\hat{\mathbf{x}}_{t|t-1})\\
\mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{h}^T\mathbf{P}_{t|t-1}\\
\hat{\mathbf{x}}_{t+1|t} &= \hat{\mathbf{F}}\hat{\mathbf{x}}_{t|t}\\
\mathbf{P}_{t+1|t} &= \hat{\mathbf{F}}\mathbf{P}_{t|t}\hat{\mathbf{F}}^T + \mathbf{G}\hat{\mathbf{Q}}\mathbf{G}^T
\end{aligned}$$

– M-Step. Estimation of the AR parameters using linear prediction. $k = 1, ..., N_s$

$$\begin{aligned}
\hat{\mathbf{s}}_{k,t} &= (\hat{\mathbf{x}}_{k,t|t})_{[1,1]}\\
\mathbf{R}_{k,t-1} &= \lambda\mathbf{R}_{k,t-2} + (1-\lambda)\mathbf{S}_k(\mathbf{x}_{t|t}\mathbf{x}_{t|t}^T + \mathbf{P}_{t|t})\mathbf{S}_k^T\\
\rho_k^{(t)} &= (\mathbf{R}_{k,t-1}^{-1})_{(1,1)}^{-1}\\
\mathbf{v}_k^{(t)} &= \rho_k\mathbf{R}_{k,t-1}^{-1}[1,0\cdots 0]^T\\
\hat{\sigma_n^2}^{(t)} &= y_t^2 - 2y_t\mathbf{h}^T\hat{\mathbf{x}}_{t|t} + \mathbf{h}^T\Big(\hat{\mathbf{x}}_{t|t}\hat{\mathbf{x}}_{t|t}^T + \mathbf{P}_{t|t}\Big)\mathbf{h}
\end{aligned}$$

The estimation of the pitches $\{T_k\}_{k=1:N_s}$ is done along with this algorithm using a multipitch estimation algorithm [12].

4 Simulations

In this section, we present some results of our source separation algorithm. We assume that the maximum number of sources is known. We limit our analysis to the case of a mixture of two simultaneous sources corrupted by white noise. In the most energetic parts of the mixture, the inconstant SNR is about 20 dB as shown in Fig. 1. We work with real speech data to which we add artificially the observation noise. The mixture consists of two voiced speech signals and is of 10 s duration. The parameters are initialized randomly, except the periods where we use a multipitch algorithm [12] running in parallel to our main algorithm. The estimated periods from the multipitch algorithm are updated in the main algorithm every 64 ms. We do two experiments. The first one is the filtering case in which all the parameters are initialized with values close to the true ones. The results are close to perfect and are shown in Fig 2. In the second experiment, the parameters are initialized randomly and estimated adaptively in the M-Step. We can see the results in Fig 3. The separation looks not very good but, when listening to the estimated sources, we find that they are under-estimated, leading to a mixture of the original sources in which the interferences are reduced. This is, in part, due to the fact that at a given moment we don't know the number of sources, so even when only one source is present, the algorithm seeks to estimate two sources. During the separation process, the estimated correlations are still polluted by the other source but the desired source is enhanced. The results can be listened on the first author personal page [1].

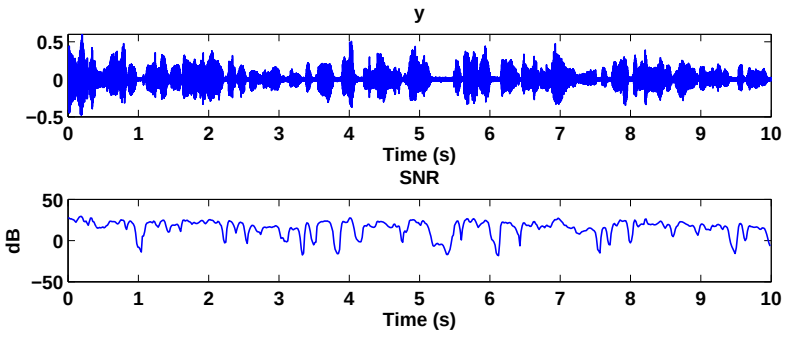

Fig. 1. Mixture and SNR evolution.

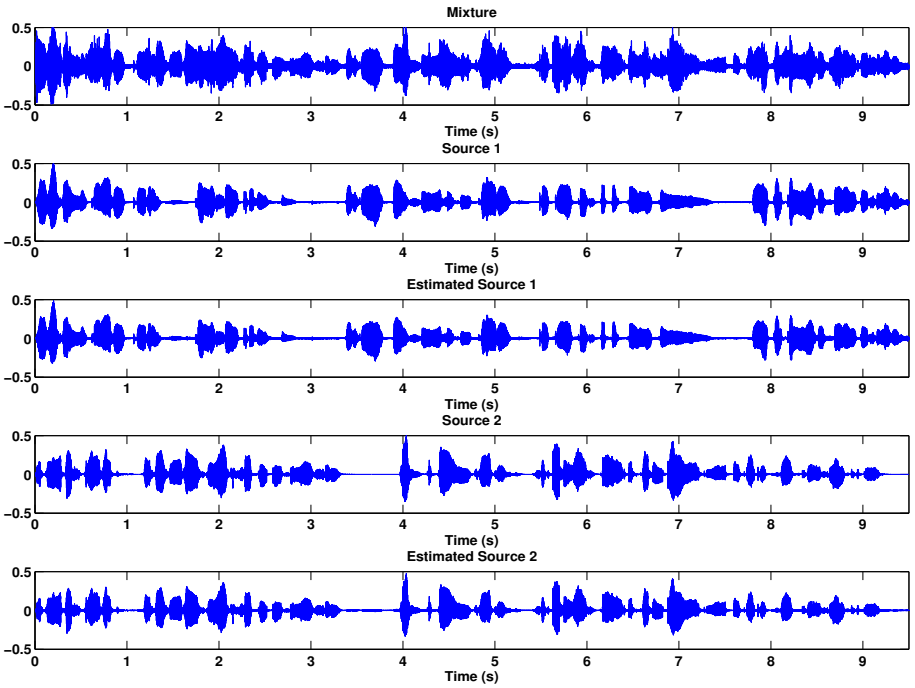

Fig. 2. Source separation with fixed and known parameters.

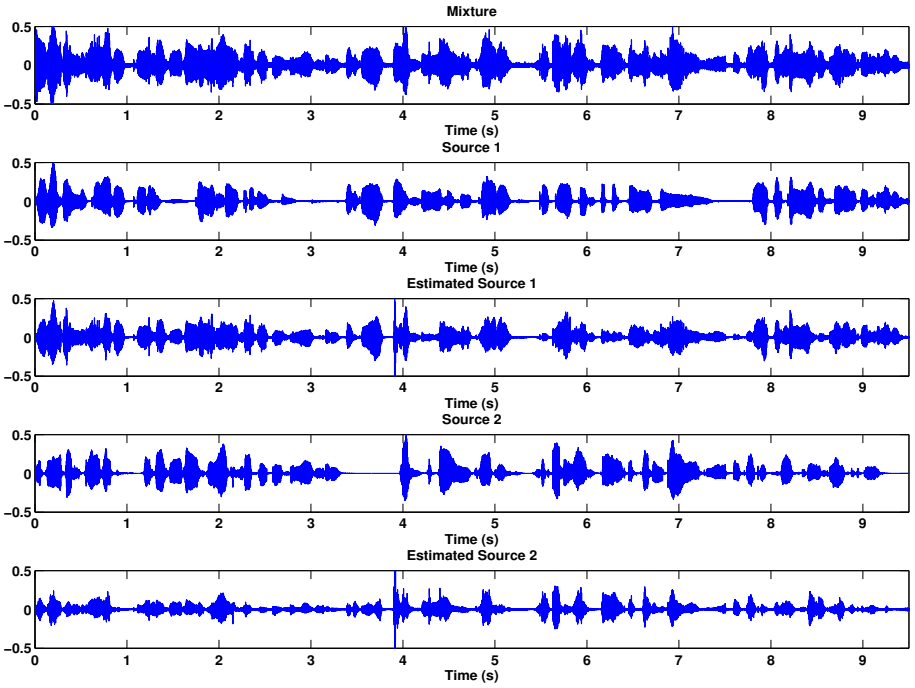

Fig. 3. Source separation adaptive estimation of the parameters.

5 Conclusion

In this paper we use the adaptive EM-Kalman algorithm for the blind audio source separation problem. The model takes into account the different aspects of speech signals production and sources are jointly estimated. The traditional smoothing step is included into the algorithm and is not an additional step. Simulations show the potential of the algorithm for real data. Yet, this performance depends a lot on the multipitch estimation quality. An error on tracking the pitches may induce the performance decreasing drastically. This work would be more complete if an other process aiming to estimate the number of active sources is working in parallel.

References

1. http://www.eurecom.fr/ ∼ bensaid/ICA10
2. A. Cichocki and R. Thawonmas, On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics, Neural Process. Lett., vol. 12, no. 1, pp. 9198, 2000.
3. A. K. Barros and A. Cichocki, Extraction of specific signals with temporal structure, Neural Comput., vol. 13, no. 9, pp. 19952003, 2001.
4. F. Tordini and F. Piazza, A semi-blind approach to the separation of real world speech mixtures, in Neural Networks, 2002. IJCNN 02. Proceedings of the 2002 International Joint Conference on, vol. 2, 2002, pp. 12931298.
5. D. Smith, J. Lukasiak, and I. Burnett, Blind speech separation using a joint model of speech production, Signal Processing Letters, IEEE, vol. 12, no. 11, pp. 784787, Nov. 2005.
6. W. C. Chu, Speech coding algorithms-foundation and evolution of standardized coders. John Wiley and Sons, NewYork, 2003.
7. M. Feder and E. Weinstein, Parameter estimation of superimposed signals using the EM algorithm, IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, pp. 477489, Apr. 1988.
8. S. Gannot, D. Burshtein, and E. Weinstein. Iterative-batch and sequential algorithms for single microphone speech enhancement. In ICASSP'98, pages 1215-8. IEEE, 1998.
9. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Society, B, 39, 1-38.
10. W. Gao, T. S., and J. Lehnert, Diversity combining for ds/ss systems with time-varying, correlated fading branches, Communications, IEEE Transactions on, vol. 51, no. 2, pp. 284295, Feb 2003.
11. C. Couvreur, Y. Bresler, Decomposition of a mixture of Gaussian AR processes, icassp, vol. 3, pp.1605-1608, Acoustics, Speech, and Signal Processing, 1995. ICASSP-95 Vol 3., 1995 International Conference on, 1995
12. Christensen, Mads and Jakobsson, Andreas and B.H., Juang,Multi-pitch estimation,Morgan & Claypool,2009