Single Microphone Blind Audio Source Separation Using Short+Long Term AR Modeling

Siouar Bensaid and Dirk Slock
EURECOM
2229 route des Crêtes, B.P. 193,
06904 Sophia Antipolis Cedex, FRANCE
Email: {siouar.bensaid, dirk.slock}@eurecom.fr

***Abstract*—In this paper, we consider the case of single microphone Blind speech separation. We exploit the joint model of speech signal (the voiced part) that consists on modeling the correlation of speech with a short term autoregressive process and its quasi-periodicity with a long term one. A linear state space model with unknown parameters is derived. The separation is achieved by estimating the state as well as the unknown parameters. This task is assured by using the Kalman filtering algorithm.**

I. Introduction

Blind Source Separation techniques are heavily needed in the speech processing domain to solve classical problems such as the 'cocktail party problem' where each speaker needs to be retrieved independently. The difficulties of speech separation can get more complex due to the impact of the propagation environment that can introduce the problem of reverberation. The description 'Blind' may not have the same impact with speech separation like it is in the general case when we do know absolutely nothing about the target sources except some hypothesis we set before such as the famous independence of sources. It is because the studies of speech signal production and modeling have revealed some distinctive features, especially the voiced part, that can be summarized in a short time correlation between samples and a quasi-periodicity introduced by the presence of pitch (fundamental frequency) of the speaker. In literature, several works considered the temporal structure of speech signal to help separation. Some work exploits only the short term correlation in speech signal and models it with a short term Auto-Regressive (AR) process [1]. Others model the quasi-periodicity of speech by introducing the fundamental frequency in the analysis [2], [3]. A last category combines the two aspects [4]and seems to get better performances. In [4], 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. In our work, we use the joint model but using only one observation. Mono-microphone case is not abundantly treated like the over-determined case or the under-determined case but with more than one observation. Some works tackled the signal microphone case but they were more likely to be classification methods based on the techniques of codebook. Since our case is relatively difficult (only a single sensor is used),we propose a rather simplified model of speech propagation : 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 to estimate the state. Since the parameters of that state space model and therefore Kalman filtering equations are unknown and should be estimated, The EM algorithm will be used for that aim ([5], [6], [7]). This paper is organized as follows: The state space model is introduced in section II. The EM-Kalman algorithm is developed in section III and the estimators' expressions are then computed. Numerical results are provided in section IV, and conclusions are drawn in section V.

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.

II. State Space Model Formulation

We consider the problem of estimating N_s mixed Gaussian sources. We use a voice production model [8], 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

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

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). In fact, on one side, it uses the short term auto-regressive model (AR) to describe the correlation between the signal samples, on the other side, it uses the long term AR model to depict the harmonic structure of speech. Let $\mathbf{x}_{k,t}$ be the vector of length (p_k+N+3), defined as $\mathbf{x}_{k,t} = [s_k(t)\ s_k(t-1)\cdots s_k(t-p_k-1) \mid \tilde{s}_k(t+1)\ \tilde{s}_k(t)\cdots\tilde{s}_k(t+1-\lfloor T_k \rfloor)\cdots\tilde{s}_k(t-N+1)]^T$. This vector can be written in terms of $\mathbf{x}_{k,t-1}$ as follows

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

where $\mathbf{g}_k$ is the (p_k+N+3) length vector defined as $\mathbf{g}_k = [\,0\,0\cdots 0 \mid 1\,0\ \cdots\cdots\ 0]^T$. The non null component is situated at the $(p_k+3)^{th}$ row. The $(p_k+N+3)\times(p_k+N+3)$ 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 $(p_k+2)\times(N+1)$ matrix $\mathbf{F}_{12,k}$ and the $(N+1)\times(N+1)$ matrix $\mathbf{F}_{22,k}$ are given by

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

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

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

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 coefficient b_k is situated in the $\lfloor T_k \rfloor$ position of the row in $\mathbf{F}_{22,k}$. 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 $N_s\times 1$ 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 $\sum_{k=1}^{N_s}(p_k+N+3)\ \times\ \sum_{k=1}^{N_s}(p_k+N+3)$ block diagonal matrix given by $\mathbf{F}$ = blockdiag $(\mathbf{F}_1,\cdots,\mathbf{F}_{N_s})$.
- $\mathbf{G}$ is the $\sum_{k=1}^{N_s}(p_k+N+3)\times N_s$ matrix given by $\mathbf{G}$ = block diag$(\mathbf{g}_1,\cdots,\mathbf{g}_{N_s})$
- $\mathbf{h}$ is the $\sum_{k=1}^{N_s}(p_k+N+3)\times 1$ 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 (p_k+N+3).

It is obvious that the linear dynamic system derived before depends on unknown parameters recapitulated in the variable $\theta = \Big\{ \{a_{k,n}\}_{k\in\{1,\ldots,N_s\}n\in\{1,\ldots,p_k\}}, \{b_k\}_{k\in\{1,\ldots,N_s\}}, \{\rho_k\}_{k\in\{1,\ldots,N_s\}}, \sigma_n^2 \Big\}$. Hence, a joint estimation of sources (the state) and θ is required. We should mention here that the pitches are considered as known. In fact, multipitch estimation is a whole issue itself where many researches have been carried and there are reliable algorithms in literature that can assure this task. In practice, before treated by our proposed algorithm the data can be first processed by a multipitch estimation algorithm in order to get the values of the pitches. In the next section, we develop the EM-Kalman of our model.

III. EM-KALMAN ALGORITHM

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 long term+short term 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 $E\left\{\breve{\mathbf{x}}_{k,t-1}\breve{\mathbf{x}}_{k,t-1}^T|y_{1:t}\right\}$. 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

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

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

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

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 $E\left\{\mathbf{x}_{k,t}\mathbf{x}_{k,t}^T|y_{1:t}\right\}$ the definition of which is

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

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. $k = 1, ..., N_s$

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

The estimation of the pitches $\{T_k\}_{k=1:N_s}$ is done along with this algorithm using a multipitch estimation algorithm [10].

IV. Numerical Results and Discussions

In the simulation part, we use artificial data similar to speech signal (artificial sources and observation noise). It consists on a noisy mixture of two sources of duration equal to $d = 128\ ms$. The pitches are respectively equal to $F_1 = 120\ Hz$ (average pitch of man voice) and $F_2 = 220\ Hz$ (average pitch of woman voice). The order of short time process is set to 5 for both. The SNR is set to 30 dB. The sampling frequency is $F_s = 16\ kHz$. In Fig. 1, we show the results of the analysis in the frequency domain. We decimate the data by factor 2 to get more visibility. The figure shows how the spectra of estimated sources are close to the original one with a little bit of distortion.

Though results with artificial data are encouraging, Simulations with real data are very critical for many reasons. The most important is the quality of estimation of the pitches. In fact, this algorithm seems to be very sensitive to pitches estimation error. An other important point is the number of sources present in the mixture. In a real context, this information is no more given like in our algorithm. Hence, if the given number of sources exceeds the real present sources, the algorithm will seek to estimate extra virtual sources.

V. 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 synthetic data. In future works, we need to improve the quality of estimation of the pitches and include a step for estimating the real number of present sources.

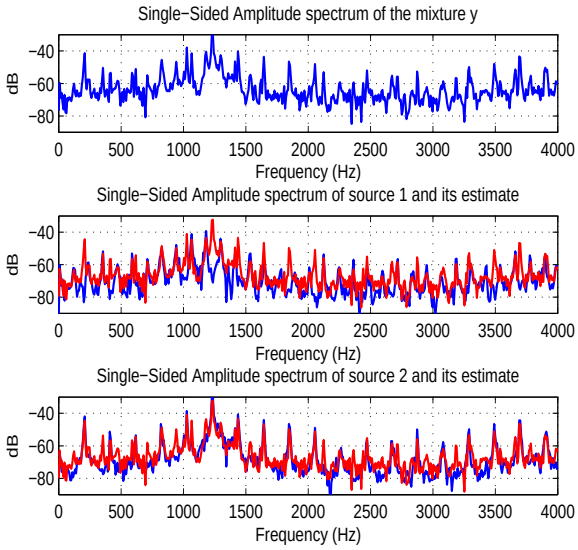

Fig. 1: Source separation : spectra of the mixture, original sources (blue) and of estimated sources (red)

REFERENCES

[1] 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. 91–98, 2000.

[2] A. K. Barros and A. Cichocki, "Extraction of specific signals with temporal structure," *Neural Comput.*, vol. 13, no. 9, pp. 1995–2003, 2001.

[3] 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. 1293–1298.

[4] 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. 784–787, Nov. 2005.

[5] C. Couvreur and Y. Bresler, "Decomposition of a mixture of gaussian ar processes," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 3, pp. 1605–1608, 1995.

[6] 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. 284–295, Feb 2003.

[7] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the em algorithm," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 4, pp. 477–489, Apr 1988.

[8] W. C. Chu, *Speech coding algorithms-foundation and evolution of standardized coders*. John Wiley and Sons, NewYork, 2003.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] M. Christensen, A. Jakobsson, and J. B.H., *Multi-pitch estimation*. Morgan & Claypool, 2009.