MONO-MICROPHONE BLIND AUDIO SOURCE SEPARATION USING EM-KALMAN FILTERS AND SHORT+LONG TERM AR MODELING

Siouar Bensaïd, Antony Schutz, Dirk Slock

Eurecom Institute , 2229 route des Crêtes, B.P. 193,
06904 Sophia Antipolis Cedex, FRANCE
e-mail:{siouar.bensaid,antony.schutz,dirk.slock}@eurecom.fr

ABSTRACT

Blind sources 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's also the more realistic one. In our approach, the autoregressive structure (short term prediction) and the periodic signature (long term prediction) of voiced speech signal are jointly modeled. The filters parameters are extracted using a combined version of the EM-Algorithm with the Rauch-Tung-Striebel optimal smoother while the fixed-lag Kalman smoother algorithm is used for the initialization.

Index Terms— Blind sources extraction, mono-microphone analysis, short+long term prediction, EM Algorithm.

1. CONTACT AUTHOR

Antony Schutz
antony.schutz@eurecom.fr.
EURECOM
Mobile Communication Department
2229 Route des Crêtes BP 193, 06904 Sophia Antipolis Cedex, France
Tel: +33 (0)4 93 00 81 98 - Fax: +33 (0)4 93 00 82 00

2. EXTENDED SUMMARY

We consider the problem of estimating an unknown number of multiple mixed Gaussian sources. We use a voice production model that can be described by filtering an excitation signal with short term prediction filter followed by a long term one and which is mathematically formulated like

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

Where y_t is the scalar observation, $x_{k,t}$ is the source k at time t. $a_{k,n}$ is the n^{th} short term coefficient of the source k while $\tilde{x}_{k,t}$ is the short term prediction error. b_k is the long term prediction coefficient of the k^{th} source, T_k its period, which is not necessary an integer, and $\{e_{k,t}\}$ are Gaussian mutually uncorrelated innovation sequences with variance ρ_k. $\{n_t\}$ is a white gaussian process with variance σ_n^2. We aim to seperate jointly these sources by estimating the short and long autoregressive (AR) coefficients of each one. We will proceed like the following: first, estimate coarsly the different pitches by extending a classical approach of monopitch estimation to the multipitch case,then use the EM algorithm to estimate the different parameters of the problem. For the first step, in speech processing it is common to use autocorrelation method (ACM) for pitch estimation, the method consists in finding the first highest peak of the function. But when several pitches are present, ACM is a bad estimator due to the fact that all the periodicities present in the signal contribute and generates proper peak and interaction peak. For estimating the different periodicity in the ACM we propose to estimate the ratio of $\frac{R_{2T}}{R_T}$ for a set of possible T (period) compared to a threshold, where R is the autocorrelation function. In order to achieve the second step (separation), we will derive a State-Space model formulated like the following

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

where $\mathbf{x}_{k,t} = [s_k(t)\ s_k(t-1) \cdots s_k(t-p_k) \mid \tilde{s}_k(t)\ \tilde{s}_k(t-1) \cdots \tilde{s}_k(t-\lfloor T_k \rfloor) \cdots \tilde{s}_k(t-N+1)]^T$
and $\mathbf{g}_k = [\,1\,0 \cdots 0 \mid 1\,0\ \cdots\cdots\ 0]^T$

$$\mathbf{F}_k = \begin{bmatrix} \mathbf{F}_{11,k} & \mathbf{F}_{12,k} \\ \mathbf{F}_{21,k} & \mathbf{F}_{22,k} \end{bmatrix}$$

EURECOM's research is partially supported by its industrial members: BMW Group Research And Technology BMW Group Company, Bouygues Telecom, Cisco Systems, France Telecom , Hitachi, SFR, Sharp, STMicroelectronics, Swisscom, Thales. The research work leading to this paper has also been partially supported by the European Commission under contract FP6-027026.

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

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

$$\mathbf{F}_{21,k} \;=\; \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{(N)\times(p_k+1)}$$

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

The N is chosen superior to the maximum value of pitches T_k in order to detect the long-term aspect. We concatenate the state equation (2) for $k \;=\; 1, \cdots, c$ and introduce an observation equation for $\{y_t\}$ to obtain the 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 $\mathbf{x}_t \;=\; [\mathbf{x}_{1,t}^T\, \mathbf{x}_{2,t}^T \cdots \mathbf{x}_{c,t}^T]^T$ and $\mathbf{e}_t = [e_{1,t}\, e_{2,t} \cdots e_{c,t}]^T$. The $(\sum_{k=1}^{c} p_k + c + Nc) \;\times\; (\sum_{k=1}^{c} p_k + c + Nc)$ block diagonal matrix $\mathbf{F}$ is given by $\mathbf{F} \;=\;$ block diag $(\mathbf{F}_1, \cdots, \mathbf{F}_c)$. The $(\sum_{k=1}^{c} p_k + c + Nc) \times c$ matrix $\mathbf{G}$ and $(\sum_{k=1}^{c} p_k + c + Nc) \times 1$ column vector $\mathbf{h}$ are given respectively by $\mathbf{G}$ = block diag$(g_1, \cdots, g_c)$, and $\mathbf{h} \;=\; [h_1^T \cdots h_c^T]^T$ where $\mathbf{h}_i \;=\; [1\, 0 \cdots 0]^T$. The covariance matrice of the input vector $\mathbf{e}_t$ is the $c \times c$ diagonal matrix $\mathbf{Q}$ =Cov$(\mathbf{e}_t)$ =diag$(\rho_1, \cdots, \rho_c)$. This State-Space model will be used in the EM algorithm. The Expectation Maximization algorithm is a general iterative method to compute ML estimates when the observed data can be regarded as "incomplete" and the incomplete data set can be related to some complete data set through a noninvertible transformation. Let the vector of observations $\mathbf{y}$ be the incomplete data set which has probability density $p(\mathbf{y}; \theta)$. The ML estimate of θ is obtained by maximizing the log-likelihood function

$$\theta_{ML} \;=\; arg\, \max_{\theta}\; log\, p(\mathbf{y}; \theta) \tag{5}$$

In our problem the data of interest are $\theta \;=\; [\rho_1 \;\cdots\; \rho_c\, \sigma_n^2\, \mathbf{v}_1 \;\cdots \mathbf{v}_c]^T$ where $\mathbf{v}_k$ contains all the short and long term coefficients of the source k. A natural choice for the complete data set $\mathcal{X}$ is the set $\mathcal{X} \;=\; [\mathbf{x}_1, \cdots, \mathbf{x}_M, \mathbf{n}_1, \cdots, \mathbf{n}_M]$. The complete data set and the incomplete data set are related in our case by the vector $\mathbf{h}$ in the equation (4). The EM Algorithm is composed of 2 steps, the expectation step (E-step) where we compute the following function
$E\; step$:

$$U(\theta, \theta^{(i)}) \;=\; E_{\theta^{(i)}}\; \{\; log\; p(\mathcal{X}; \theta) \;\mid\; \mathbf{y}\; \} \tag{6}$$

and the maximization step where we maximize (6)
$M\; step$:

$$\theta^{(i+1)} \;=\; arg_\theta\; U(\theta, \theta^{(i)}) \tag{7}$$

here i denote the iteration indice. In order to compute (6), we first evaluate the log-likelihood function $log\; p(\mathcal{X}; \theta)$:

$$\begin{aligned} \log\; p(\mathcal{X}; \theta) &= \sum_{k=1}^{c} log\; p(\mathbf{x}_{k,1}, \cdots, \mathbf{x}_{k,M}) \;+\; log\; p(n_1, \cdots, n_M) \\ &= \sum_{k=1}^{c} (\; log\; p(\mathbf{e}_{k,2}, \cdots, \mathbf{x}_{k,M} \;\mid\; \mathbf{x}_{k,1}) \;+\; log\; p(\mathbf{x}_{k,1})\;) \\ &\qquad + log\; p(n_1, \cdots, n_M) \end{aligned}$$

Using the Gaussian assumption, we find, after some algebraic manipulation

$$\begin{aligned} &\log\; p(\mathcal{X}; \theta) = \textstyle\sum_{k=1}^{c} \Big[-\frac{1}{2}\, (M + N + p_k)\; log\; \rho_k \\ &-\frac{1}{2\,\rho_k} \left(\mathbf{v}_k^T \left(\sum_{t=2}^{M} \breve{\mathbf{x}}_{k,t}\, \breve{\mathbf{x}}_{k,t}^T \right) \mathbf{v}_k \;+\; trace\; \{ \Gamma_k^{-1}\, \mathbf{x}_{k,1}\, \mathbf{x}_{k,1} \} \right) \\ &-\frac{M}{2}\; log\; \sigma_n^2 \;-\; \frac{1}{2\,\sigma_n^2} \sum_{t=1}^{M} \left(y_t^2 \;-\; 2\, y_t\, \mathbf{h}^T\, \mathbf{x}_t \;+\; \mathbf{h}^T\, \mathbf{x}_t\, \mathbf{x}_t^T\, \mathbf{h} \right) \\ &+ K \end{aligned}$$

where $\breve{\mathbf{x}}_{k,t} \;=\; [s_k(t, \theta) \cdots s_k(t - p_k, \theta)\; \tilde{s}_k(t - \lfloor T_k \rfloor, \theta)\; \tilde{s}_k(t - \lfloor T_k \rfloor - 1, \theta)]^T \;=\; \mathbf{S}_k^T\, \mathbf{x}_{k,t}$, $\mathbf{S}_k^T$ is a transformation matrix which extract the partial state $\breve{\mathbf{x}}_{k,t}$ from the complete state $\mathbf{x}_{k,t}$, $\mathbf{v}_k \;=\; [1 \;-\; a_{k,1} \cdots - a_{k,p_k} \;-\; \alpha\; b_k \;-\; (1-\alpha)\; b_k]^T$, K is a constant independent of θ and Γ_k is the normalized covariance matrix $\Gamma_k \;=\; \rho_k^{-1}\; \mathrm{Cov}(\mathbf{x}_{k,1})$.After applying the $E_{\theta^{(i)}}\; \{.\;\mid\; \mathbf{y}\; \}$ operator we deduce that

$$\begin{aligned} &U(\theta, \theta^{(i)}) \;=\; \textstyle\sum_{k=1}^{c} \Big[-\frac{1}{2}(M + N + p_k)\; log\; \rho_k \\ &-\frac{1}{2\rho_k} \left(\mathbf{v}_k^T\, \mathbf{S}_k^T \left(\sum_{t=2}^{M} \left(\hat{\mathbf{x}}_{k,t|M}^{(i)} \left(\hat{\mathbf{x}}_{k,t|M}^{(i)} \right)^T + \mathbf{P}_{k,t|M}^{(i)} \right) \right) \mathbf{S}_k\, \mathbf{v}_k + \; trace \left\{ \Gamma_k^{-1} \left(\hat{\mathbf{x}}_{k,1|M}^{(i)}\, (\hat{\mathbf{x}}_{k,1|M}^{(i)})^T \right. \right. \right. \\ &\left. \left. \left. + \mathbf{P}_{k,1|M}^{(i)} \right) \right\} \right) \Big] - \frac{M}{2}\; log\; \sigma_n^2 \;-\; \frac{1}{2\sigma_n^2} \sum_{t=1}^{M} \left(y_t^2 \right. \\ &\left. -2 y_t \mathbf{h}^T \hat{\mathbf{x}}_{t|M}^{(i)} + \mathbf{h}^T \left(\hat{\mathbf{x}}_{t|M}^{(i)} \left(\hat{\mathbf{x}}_{t|M}^{(i)} \right)^T \;+\; \mathbf{P}_{t|M}^{(i)} \right) \mathbf{h} \right) \;+ K \end{aligned} \tag{8}$$

where $\hat{\mathbf{x}}_{t|M}^{(i)} = E_{\theta^{(k)}}\ \{\mathbf{x}_t \mid \mathbf{y}\}$ and $\hat{\mathbf{x}}_{k,t|M}^{(i)} = E_{\theta^{(k)}}\ \{\mathbf{x}_{k,t} \mid \mathbf{y}\}$ denote the smoothed conditional means, $\mathbf{P}_{t|M}^{(i)} = Cov_{\theta^{(k)}}\ (\mathbf{x}_t \mid \mathbf{y})$ and $\mathbf{P}_{k,t|M}^{(i)} = Cov_{\theta^{(k)}}\ (\mathbf{x}_{k,t} \mid \mathbf{y})$ are the smoothed conditional covariance matrices. These quantities are calculated in the E-step using the Rauch-Tung-Striebel optimal smoother like described below

Set $\hat{\mathbf{x}}_{1|0} = 0$
Set $\mathbf{P}_{1|0} = block\ diag(\rho_1\ \Gamma_1, \cdots, \rho_c \Gamma_c)$
$For\ t = 1\ to\ M\ do$
$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{h}(\mathbf{h}^T\mathbf{P}_{t|t-1}\mathbf{h} + \sigma_n^2)^{-1}$
$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{h}^T\hat{\mathbf{x}}_{t|t-1})$
$\hat{\mathbf{x}}_{t+1|t} = \hat{\mathbf{F}}\hat{\mathbf{x}}_{t|t}$
$\mathbf{P}_{t+1|t} = \hat{\mathbf{F}}\mathbf{P}_{t|t}\hat{\mathbf{F}}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T$
$For\ t = M\ to\ 1\ do$
$\mathbf{A}_t = \mathbf{P}_{t|t}\hat{\mathbf{F}}^T\mathbf{P}_{t+1|t}^{-1}$
$\hat{\mathbf{x}}_{t|M} = \hat{\mathbf{x}}_{t|t} + \mathbf{A}_t(\hat{\mathbf{x}}_{t+1|M} - \hat{\mathbf{F}}\hat{\mathbf{x}}_{t|t})$
$\mathbf{P}_{t|M} = \mathbf{P}_{t|t} + \mathbf{A}_t(\mathbf{P}_{t+1|M} - \mathbf{P}_{t+1|t})\mathbf{A}_t^T$

The maximization of (8) during the M-Step is trivial. Setting the partial derivative with respect to $\rho_1, \cdots, \rho_c,\ \sigma_n^2$ to zero yields to the new estimates of this variables. To estimate the vectors $\mathbf{v}_k$ we use the fact that $\mathbf{e}_{k,t} = \breve{\mathbf{x}}_{k,t}^T\mathbf{v}_k$, by multiplying on the left with $\breve{\mathbf{x}}_{k,t}$ and applying the operator $E_{\theta^{(i)}}\ \{. \mid \mathbf{y}\}$ we will obtain

$$\mathbf{S}_k^T\ \frac{1}{M}\left(\ \sum_{t=1}^{M}\ \left(\hat{\mathbf{x}}_{k,t|M}^{(i)}\ \left(\hat{\mathbf{x}}_{k,t|M}^{(i)}\right)^T + \mathbf{P}_{k,t|M}^{(i)}\right)\right)\ \mathbf{S}_k\ \mathbf{v}_k = [\rho_k\ 0\ \cdots\ \rho_k\ 0\ \cdots\ \cdots\ 0]^T$$

where the second ρ_k is the $(p_k + 2)^{th}$ component. $\mathbf{F}$ depends of the parameters of the problem, therefore, it should be computed in each iteration in order to be used in the Kalman algorithm (hence its estimate $\hat{\mathbf{F}}$). Converging to pretty good values of parameters requires a good initialization. Therefore, here we will proceed to a first sweep where we use the fixed-lag Kalman smoothing algorithm. The obtained values will be used as initializations to the the Rauch-Tung-Striebel optimal smoother algorithm described above and that will be runned several times till convergence.