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 the use of Kalman filtering modified with the variational bayes techniques which takes into consideration the estimation error of parameters used in Kalman filter.

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 independency 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 works exploit 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 ascendent 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 the 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.

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

\[
\begin{align*}
  y_t &= \sum_{k=1}^{N_s} s_{k,t} + n_t, \\
  s_{k,t} &= \sum_{n=1}^{p_k} a_{k,n} s_{k,t-n} + \tilde{s}_{k,t} \\
  \tilde{s}_{k,t} &= b_k \hat{s}_{k,t-T_k} + e_{k,t} 
\end{align*}
\]

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.
\( \{ e_{k,t} \}_{k=1}^{c} \) are the independent Gaussian distributed innovation sequences with variance \( \rho_k \).

* \( \{ n_t \} \) is a white Gaussian process with variance \( \sigma_n^2 \), 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) because, on one side, it uses the short term autoregressive 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 \( x_{k,t} \) be the vector of length \((p_k + N + 2)\), defined like \( x_{k,t} = [s_k(t) s_k(t-1) \cdots s_k(t-p_k-1) \, \hat{s}_k(t+1) \cdots \hat{s}_k(t-N+1)]^T \). This vector can be written in terms of \( x_{k,t-1} \) as the following

\[
x_{k,t} = F_k x_{k,t-1} + g_k e_{k,t} \tag{2}
\]

where \( g_k \) is the \((p_k + N + 2)\) length vector defined as \( g_k = [0 \, 0 \, \cdots 0 \, | \, 1 \, 0 \, \cdots 0]^T \). The non null component is at the \((p_k + 3)\) position. The \((p_k + N + 2) \times (p_k + N + 2)\) matrix \( F_k \) has the following structure

\[
F_k = \begin{bmatrix}
F_{11,k} & F_{12,k} \\
F_{12,k} & O & F_{22,k}
\end{bmatrix}
\]

where the \((p + 2) \times (p + 2)\) matrix \( F_{11,k} \), the \((p + 2) \times N\) matrix \( F_{12,k} \) and the \( N \times N\) matrix \( F_{22,k} \) are given by

\[
F_{11,k} = \begin{bmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 \\
\end{bmatrix}
\]

\[
F_{12,k} = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}
\]

\[
F_{22,k} = \begin{bmatrix}
0 & 0 & b_k & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\]

It is noteworthy that the choice of the \( F_{22,k} \) 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 \([T_k]\) position of the row in \( F_{22,k} \). Since \( N_s \) sources are present, we introduce the vector \( x_t \) that consists of the concatenation of the \( \{ x_{k,t} \}_{k=1}^{N_s} \) vectors \( x_t = [x_{1,t}^T x_{2,t}^T \cdots x_{N_s,t}^T]^T \) which results in the vector difference equation 3. Moreover, by reformulating the expression of \( \{ y_t \} \), we introduce the observation equation 4.

We obtain the following state space model

\[
x_t = F x_{t-1} + G e_t \tag{3}
\]

\[
y_t = h^T x_t + n_t \tag{4}
\]

where

\[
e_t = [e_{1,t} e_{2,t} \cdots e_{N_s,t}]^T \text{ is the } N_s \times 1 \text{ column vector resulting of the concatenation of the } N_s \text{ innovations. Its covariance matrix is the } N_s \times N_s \text{ diagonal matrix } Q = \text{diag}(\rho_{11}, \ldots, \rho_{N_sN_s}).
\]

\[
F \text{ is the } \sum_{k=1}^{N_s} (p_k + N + 2) \times \sum_{k=1}^{N_s} (p_k + N + 2) \text{ block diagonal matrix given by } F = \text{ blockdiag}(F_1, \ldots, F_{N_s}).
\]

\[
G \text{ is the } \sum_{k=1}^{N_s} (p_k + |T_k| + 2) \times N_s \text{ matrix given by } G = \text{ block diag}(g_1, \ldots, g_{N_s}).
\]

\[
h \text{ is the } \sum_{k=1}^{N_s} (p_k + |T_k| + 2) \times 1 \text{ column vector given by } h = [h_1^T \cdots h_{N_s}^T]^T \text{ where } h_i = [1 \, 0 \, \cdots 0]^T \text{ of length } (p_k + N + 2).
\]

It is obvious that the linear dynamic system derived before depends on unknown parameters recapitulated in the variable \( \theta = \{ \{ a_{k,n} \}_{k=1}^{1} \ldots, a_{k,n} \}_{n=1}^{N_s}, \{ b_k \}_{k=1}^{N_s}, \{ p_k \}_{k=1}^{N_s}, \sigma_n \}. \) Hence, a joint estimation of sources (the state) and \( \theta \) is required. We could mention here that the pitches are considered as known. Infact, multipitch estimation is a whole issue itself where many researches have been carried and there are reliable algorithms in litterature 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. In the M-step, an estimate \( \hat{\theta} \) is computed. 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 LT+ST coefficients could be written like \( e_{k,t-1} = v_k x_{k,t-1} \), where \( v_k = [1 \, -a_{k,1} \, \cdots \, -a_{k,p_k} \, -b_k]^T \) is \((p_k+2) \times 1\) column vector and \( x_{k,t-1} = [s_k(t-1, \theta) \, \cdots \, s_k(t-(1-p_k, \theta) \, \hat{s}_k(t+1, \theta) \, \cdots \, \hat{s}_k(t-(N-1, \theta) \, \hat{s}_k(t-N+1, \theta)]^T \) is called the partial state since it is not the complet \( x_{k,t-1} \). After some mathematical manipulation, the following relation between the vector of coefficients and the innovation power is deduced

\[
v_k = \rho_k R_{k,t-1}^{-1} [1, 0 \cdots 0]^T \tag{5}
\]

where the covariance matrix \( R_{k,t-1}^{-1} \) is defined as \( E \{ x_{k,t-1} x_{k,t-1}^T \} \). It is important to notice that the estimation of \( \hat{x}_{k,t-1} x_{k,t-1}^T \) is done using observations till time \( t \), which coincides with a fixed-lag smoothing treatment \((lag = 1)\). Another key relation, between the partial state at time \( t-1 \) and the full state at time \( t \) is

\[
\hat{x}_{k,t-1} = S_k x_{k,t} \text{ where } S_k \text{ is a selection matrix. This can be used in the partial state covariance matrix computing to get } R_{k,t-1}^{-1} = S_k E \{ x_{k,t} x_{k,t}^T | y_{t,t} \} S_k^T.
\]
transition from the fixed lag smoothing with the partial state to the simple filtering with the full state, which justifies the augmented form matrix $F_k$ or more precisely $F_{11,k}$. The innovation power is simply deduced from $R_{k,t-1}$ as the first component of the matrix $R_{k,t-1}$. This latter is computed in the Estep using the classic Kalman filter equations.

The adaptive algorithm is presented as Algorithm 1. The algorithm needs an accurate initialization, which will be discussed afterwards. In the algorithm $\hat{s}_{k,t}$ is the estimation of the source $k$ at time $t$ and $P_{k}$ is the $k^{th}$ diagonal block of $P_{t|t}$.

### Adaptive EM Kalman Algorithm

- **E-Step.** Estimation of the sources covariance

  \[
  K_t = P_{t|t-1} h^T P_{t|t-1} h + \sigma_n^2
  \]

  \[
  \hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - h^T \hat{x}_{t|t-1})
  \]

  \[
  P_{t|t} = P_{t|t-1} - K_t h^T P_{t|t-1}
  \]

  \[
  \hat{P}_{x_{t+1}} = \hat{F} \hat{P}_{x_{t}} \hat{F}^T + G Q G^T
  \]

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

  \[
  \hat{s}_{k,t} = (\hat{x}_{t|t}^T h (c-1) (N+p+2))
  \]

  \[
  \hat{P}_k = S_k P_k S_k^T
  \]

  \[
  R_k = \lambda R_k + (1-\lambda) (\hat{s}_{k} \hat{s}_{k}^T + \hat{P}_k)
  \]

  \[
  \rho_k = (R_k^{-1})^{(-1)}_{11}
  \]

  \[
  \nu_{k,t} = \rho_k R_k^{-1}
  \]

### IV. Numerical Results

In this section we show some results for the source separation problem. We assume the number of sources to be known and we limit our analysis to two simultaneous sources corrupted by white noise. We investigate three cases, the first one assume to know perfectly all the parameters for the initialization. For the second one, the parameters are estimated on the observation, the third case is the filtering case and will be used for the comparison.

For estimating the parameters on the observation a multipitch like algorithm is used, when we know the period of each sources we construct a correlation sequence with the value taken in the spectrum, the long term coefficient and the variance of the excitation of each source are taken equal to one.

For evaluating the performance of our BASS algorithm we present some results on SNR estimation and the MSE. For each value of the SNR 1000 signals are generated using random parameters. The periods varies between $80Hz$ and $320Hz$ and are not necessarily integer, between [01] for the long term coefficient and the variance of the excitation. The short term coefficients of order 10 are Gaussian but constrained to give reasonable source (stable).

Fig. 1 and Fig. 3 shows the results of the analysis. The estimated parameters initialization results are very closed to the exact initialisation. The two initialization converged rapidly to the filtered solution which is the best solution we can obtain with our model. We show in Fig. 2 and Fig. 4 the obtained waveforms and spectra. We can show that the spectral shape and the harmonic structure are well modeled.

### 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 real and synthetic data. In future works, we intend to use the Variational Bayes framework (VB) instead of the EM framework. In fact, the VB algorithm is a generalization of the EM. Since the former considers the parameters like random and hence compute their estimation errors and use them to correct the estimation iteratively, while the latter uses only their expectation and neglects the error estimation impact ([9], [10], [11], [12], [13], [14]).

REFERENCES