Blind Audio Source Separation using Short+Long Term AR Source Models and Iterative Itakura-Saito Distance Minimization

Antony Schutz and Dirk Slock
EURECOM
Mobile Communications Dept.
2229 Route des Crêtes, BP 193, 06904 Sophia Antipolis Cedex, France
Email: {antony.schutz, dirk.slock}@eurecom.fr

***Abstract*—Blind audio source separation (BASS) arises in a number of applications in speech and music processing such as speech enhancement, speaker diarization, automated music transcription etc. Generally, BASS methods consider multichannel signal capture. The single microphone case is the most difficult underdetermined case, but it often arises in practice. In the approach considered here, the main source identifiability comes from exploiting the presumed quasi-periodic nature of sources via long-term autoregressive (AR) modeling. Indeed, musical note signals are quasi-periodic and so is voiced speech, which constitutes the most energetic part of speech signals. We furthermore exploit (e.g. speaker or instrument related) prior information in the spectral envelope of the source signals via short-term AR modeling. We present an iterative method based on the minimization of the Itakura-Saito distance for estimating the sources parameters directly from the mixture using a frame based analysis.**

I. Introduction

The need for Blind Audio Source Separation (BASS) arises with various real-world signals, including speech enhancement, speaker diarization, automated music transcription etc.. Generally, BASS methods consider multichannel signal capture and has been dealt with extensively in the literature. In the over determined case of BSS the source separation can be performed satisfactorily, especially in clean environment, for example by using Independent Component Analysis (ICA) [1], [2] or Computational Auditory Scene Analysis (CASA) [3]. For underdetermined BSS (UBSS), the problem is ill-defined and its solution requires some additional assumptions. In the approach considered here, the sound is modeled as a sum of autoregressive processes with an additive white noise. Each source is assumed to have a periodic nature wich makes its presence identifiable, in [4] we have presented a separation algorithm which gives good results when the parameters are well estimated. Here, the parameters are estimated from the mixture, assuming the number of sources known, by extracting the sources correlation instead of using the separated sources.

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.

This paper is organized as follow. In section II we present the model of joint speech production. In section III and IV we present the method for estimating the parameters. Then, in section VI, we give some results.

II. Model

A. Signal Model

We consider the problem of estimating an unknown number of multiple mixed Gaussian sources. We use the short+long term autoregressive (AR) voice production model [5]:

$$y_t = \sum_{k=1}^{K} x_{k,t} + v_t, \tag{1}$$

$$x_{k,t} = \sum_{n=1}^{p_k} a_{k,n}\, x_{k,t-n} + \tilde{x}_{k,t} \tag{2}$$

$$\tilde{x}_{k,t} = b_k\, \tilde{x}_{k,t-\tau_k} + e_{k,t} \tag{3}$$

Such source models are frequently used in voice encoding algorithms like CELP and LPC. Here, y_t is the measured mixture of signals, K is the number of sources x_k. v_t is an additive white gaussian noise of variance σ_v^2 and is supposed to be uncorrelated with the sources. $e_{k,t}$ is the excitation signal of the source k also assumed to be gaussian with variance σ_k^2. For each source x_k, τ_k is the period (its fractional part can be implemented by linear interpolation if the sampling frequency is high enough), b_k its long-term prediction coefficient and the short-term prediction coefficients, of order p_k, are $a_{k,n}$.

If we introduce the short-term and long-term prediction error transfer functions

$$A_k(z) = \sum_{n=0}^{p_k} a_{k,n}\, z^{-n} \tag{4}$$

$$B_k(z) = 1 - b_k\, z^{-\tau_k} \tag{5}$$

with $a_{k,0} = 1$, the spectra of the sources can be written as:

$$S_k(f) = \frac{\sigma_k^2}{|A_k(f)\, B_k(f)|^2} \tag{6}$$

$$S_0(f) = \sigma_v^2 \tag{7}$$

The additive noise is considered as an AR model of order 0 and is included in the signals set.

The sources separation algorithm is based on the assumption that the sources can be extracted from the mixture using the knowledge of the parameters, this implies a good estimation of the related parameters.

B. parameters subsets

If the parameters can be considered constant during a short time segment we can use a frame based method. The short and long term aspect of the signals are very different by nature, it may seem natural to separating their analysis. Except the additive noise, the parameters are sources related, we group them by source; this impose to alternate the estimation of a group between sources. The overall set of parameters contains the following subsets (short term and long term parameters):

$$\theta = [\theta_1^T \cdots \theta_k^T \ \sigma_v^2]^T \quad (8)$$
$$\theta_k = [\ \mathbf{a}_k \ \varphi_{\mathbf{k}}]^{\mathbf{T}} \quad (9)$$
$$\mathbf{a}_k = [a_{k,1} \cdots a_{k,p_k}]^T \quad (10)$$
$$\varphi_{\mathbf{k}} = [\ b_k \ \tau_k \ \sigma_k^2 \ ]^T \quad (11)$$

For the estimation of a given subset of parameters of a given source we consider that the other sources are constant and also the other subset of the current source.

III. Parameters Estimation

Many approach can be used for estimating the AR coefficients from a mixture. Here we propose to minimize the Itakura-Saito (IS) distance, then, most of the derivation are done in the spectral domain and need low complexity operations. Consider the IS distance:

$$\int df[ln(x) - x + 1] \quad (12)$$
$$x = \frac{S'(f)}{S(f;\theta)} \quad (13)$$

Where $S'(f) = |y(f)|^2$ is the sample spectrum and $S(f;\theta) = \sum_k S_k(f;\theta_k)$ with $S_k(f;\theta_k)$ the parametric spectrum of the source k, defined in (6).

Close to the solution $x \approx 1$ and so $ln(x) - x + 1 \approx \frac{1}{2}(x-1)^2 + O((x-1)^3)$. Updating the parameters of the source k while keeping the other sources parameters constant, we can rewrite x as:

$$\frac{S'(f)}{S(f;\theta)} = \frac{S'(f)}{S_k(f;\theta_k)} \ \frac{1}{1 + S_k(f;\theta_k)^{-1} \sum_{\bar{\mathbf{k}} \neq k} S_{\bar{\mathbf{k}}}(f;\theta_{\bar{\mathbf{k}}})} \quad (14)$$

The minimization of the IS distance compared to $S_k(\theta_k)$ leads to a linear prediction problem on the following spectrum:

$$\frac{S'(f)}{S(f;\theta)} = \frac{y(f)}{1 + S_k(f;\theta_k)^{-1} \sum_{\bar{\mathbf{k}} \neq k} S_{\bar{\mathbf{k}}}(f;\theta_{\bar{\mathbf{k}}})} y^*(f) \quad (15)$$

which is the crossed spectrum between the Wiener estimates of the source k and the observation y.

IV. Algorithm

The algorithm consists by alternatively estimating the short term and long term subsets of parameters. Each subsets estimation needs to be iterated between all the sources (including the additive noise) until convergence. Also, we iterate between all the sources and the algorithm is stopped when all the subsets of all sources have converged. The short and long term parameters are in the spectral domain:

$$A_k = F\ [1 \ -\mathbf{a}_k \ 0 \cdots 0]^T \quad (16)$$
$$B_k = F\ [1 \ 0 \cdots - (1-\alpha_k) b_k \ -\alpha_k b_k \ 0 \cdots 0]^T \quad (17)$$

Where F is the DFT Matrix and the two vectors are zero padded, A_k and B_k have the same size (N) as the DFT of the observation. α_k is the interpolation coefficient of the source k, due to non integer period. The two terms in b_k are at the position $\lfloor \tau_k \rfloor$ and $\lfloor \tau_k + 1 \rfloor$.

$S_k(f;\theta_k)$ and $S_0(f)$ are defined in (6) and (7) respectively. For convenience, we also define

$$S_{\bar{\mathbf{k}}}(f;\theta_{\bar{\mathbf{k}}}) = \sum_{\bar{\mathbf{k}}} \frac{\sigma_{\bar{\mathbf{k}}}^2}{|A_{\bar{\mathbf{k}}}(f) B_{\bar{\mathbf{k}}}(f)|^2} \quad (18)$$

The index $\bar{\mathbf{k}}$ include all the sources (and the noise) except the one of interest, the source k.

A. Short term parameters estimation

For estimating the short term (st) parameters of order $p+1$, also if we work with a single source, we have to remove the effect of the long term otherwise the estimation is biased by the harmonic structure. So, until convergence, and for all the sources:

$$\hat{S}_k(f) = \frac{S'(f)}{1 + S_k^{-1}(f;\theta_k) S_{\bar{\mathbf{k}}}(f;\theta_{\bar{\mathbf{k}}})} \quad (19)$$
$$S_k^{st}(f) = \hat{S}_k(f) \ |B_k(f)|^2 \quad (20)$$
$$r_k = F^{-1} \ S_k^{st}(f) \quad (21)$$

The short term coefficients are computed on the above correlation sequence and the new estimates of $\mathbf{a}_k$ is used for the next source.

B. Long term parameters estimation

The long term (lt) parameters consists on three parameters, we also need to clean the short term influence for estimating them. So, until convergence, and for all the sources:

$$\hat{S}_k(f) = \frac{S'(f)}{1 + S_k^{-1}(f;\theta_k) S_{\bar{\mathbf{k}}}(f;\theta_{\bar{\mathbf{k}}})} \quad (22)$$
$$S_k^{lt}(f) = \hat{S}_k(f) \ |A_k(f)|^2 \quad (23)$$
$$r_k = F^{-1} \ S_k^{lt}(f) \quad (24)$$

If the period is known $b_k = \frac{r_k(\tau_k)}{r_k(0)}$ otherwise τ_k is estimated as the delay wich maximize b_k (for a realistic range of delay).

The short+plus long term prediction error is:

$$S^e(f) = S_k^{lt}(f)\ |B_k(f)|^2 \quad (25)$$

$$\sigma_k^2 = \frac{1}{N}\sum_f S^e(f) \quad (26)$$

C. Noise variance

In the above formulation the noise is treated as a source. It is the only one global parameter (non related to a particular source) and needs the knowledge of all the sources parameters.

$$S_0(f) = \frac{S'(f)}{1 + S_0^{-1}(f)\sum_{k=1} S_k(f;\theta_k)} \quad (27)$$

$$\sigma_v^2 = \frac{1}{N}\sum_f S_0(f) \quad (28)$$

V. Initialization and details

The algorithm needs to be initialized. The periods (for voiced speech segment) are estimated using a multipitch algorithm, see for example [6], and are adapted during the iterative process. For the short and the long term estimation, the initialization must be more precise, ifnot, some holes may appear in the estimated spectrum leading to instability. A way for avoiding this situation, and independently of the initialization, is to control the variation of the correlation sequence used in (21) and (24) using forgetting factors. Like this, correlations are slowly adapted and allow a kind of tracking for multiframe processing. But it still needs to be initialized, for a single frame we choose for the short and the long term coefficients to set them to zeros. At the begining of the algorithm this is equivalent of constructing the correlation sequence (21) with the spectral peaks information (the most energetic part of the spectrum) for the short term and to estimate the long term coefficient on the correlation sequence (24) in the mixture, but, step by step this leads to estimate the parameters on the correlation sequence of the source individually. When working on long speech segments we have several consecutive frames. The initialization of a new frame analysis is done using estimated parameters and correlation sequences from the previous frame. When the energy of the observation becomes too small, we stop the analysis and when a signal reappears we re-initialize everything.

VI. Simulation

A. Synthetic data

The first simulation consists on applying the algorithm on a completly synthetic spectrum, define as :

$$Y(f) = \sum S_k(f) + \sigma_n^2 \quad (29)$$

With $S_k(f)$ defined in (6). The result is shown on Fig 1. As the signal is synthetic, and corresponds to the model, the result is almost perfect. For this example all the noises variances are equal to one and the periods are integer. The estimated short term coefficients are presented in table I.

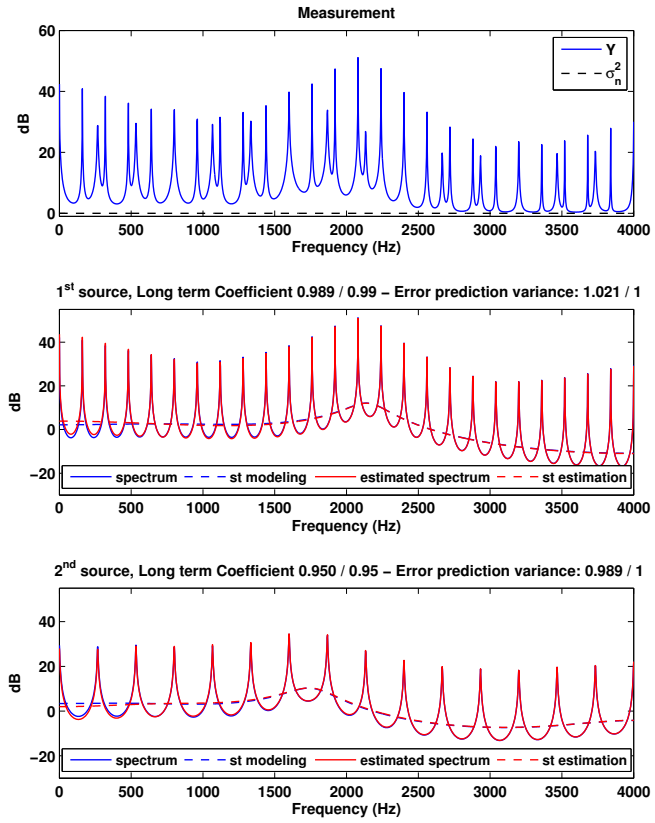

Fig. 1. Spectrogram of the mixture, sources, estimated sources and associated short term modeling.

TABLE I
Short term coefficients estimation.

$a_{k=1,n}$	0.6900	-0.9023	0.6635	-0.2310	0.0025
$\hat{a}_{k=1,n}$	0.7220	-0.8818	0.7140	-0.2154	0.0192
$a_{k=2,n}$	0.7113	-0.5150	-0.0655	0.3670	-0.1737
$\hat{a}_{k=2,n}$	0.6847	-0.5407	-0.0866	0.3393	-0.1888

B. Real Speech Segment

The next simulation song is composed of two english speakers segment, a man and a woman. The lenght of the segment is 64 ms at 8 KHz, the mixture is artificially made and the signal to noise ratio (SNR) is fixed to 20 dB, the periods are estimated using a multipitch estimator. We use the output of the presented algorithm for making the separation, we compare the obtained sources to the original sources and the sources extracted using the parameters estimated on the individual sources (before the mixing process). The separation algorithm, presented in [4], extracts windowed sources so the estimated sources are also windowed. The waveform of the decomposition is shown on Fig 2. The difference between the two extracted sources is low, note that the speech segments are well voiced. In Fig 3 we show the associated spectra, as we can see the spectra of the extracted sources with the proposed method is not as closed to the true one as the one obtained with the parameters estimated on the sources, but the most energetic part is correctly modeled.

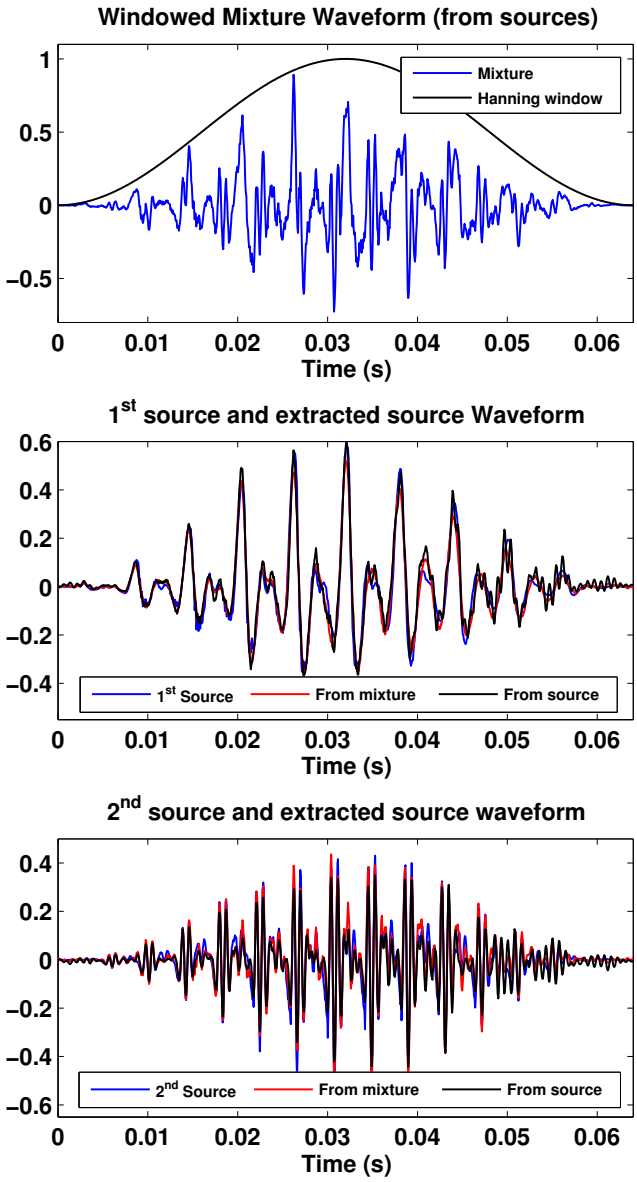

Fig. 2. Waveform of the mixture, sources and estimated sources. The sources are extracted with the parameters from the mixture and from the source.

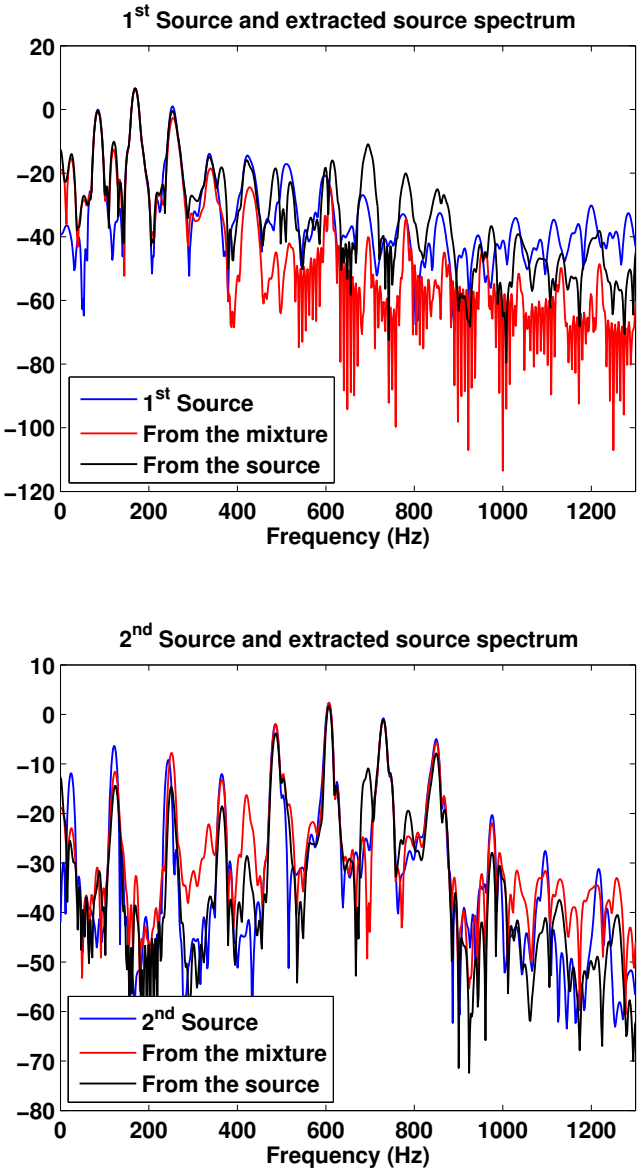

Fig. 3. Spectrum of the mixture, sources and estimated sources. The sources are extracted with the parameters from the mixture and from the source.

VII. Conclusion

In this paper we have proposed an algorithm, based on the minimization of the Itakura-Saito distance, for estimating the short+long term AR parameters of several sources and also the additive noise variance from a mixture. The AR parameters are estimated on the extracted sources correlations which doesn't need to extracts the sources themself, only the spectrum of the sources are approximated. The algorithm is iterative and need a robust initialization which, also, has been presented. Simulations on synthetic and real data are encouraging. The estimated parameters leads to a separation result which is closed to the one obtained by using the parameters estimated on the individual sources.

References

[1] A.Hyvarinen, "Survey on independent component analysis," *Neural Computing Surveys, Vol. 2, pp. 94-128*, 1999.

[2] M. Casey, "Separation of mixed audio sources by independent subspaces analysis," *int. Computer Music Conference, Berlin, August*, 2000.

[3] D. Rosenthal and H. Okuno, "Computational auditory scene analysis," *LEA Publishers, Mahwah NJ*, 1998.

[4] A. Schutz and D. T. M. Slock, "Single-microphone blind audio source separation via Gaussian Short+Long Term AR Models," in *ISCCSP 2010, 4th International Symposium on Communications, Control and Signal Processing, March 3-5, 2010, Limassol, Cyprus*, 03 2010.

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

[6] M. G. Christensen, P. Stoica, A. Jakobsson, and S. Holdt Jensen, "Multi-pitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, 2008.