BLIND AUDIO SOURCE SEPARATION USING SHORT+LONG TERM AR SOURCE MODELS AND SPECTRUM MATCHING

Antony Schutz, 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 the 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 (weighted) Itakura-Saito distance for estimating the source parameters directly from the mixture using frame based processing.

1. 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. The topic has been dealt with extensively in the literature. In the (over) determined case the source separation can be performed satisfactorily, especially in a clean environment, for example by using Independent Component Analysis (ICA) [1, 2]. For underdetermined BSS (UBSS) and especially for the single sensor case considered here, the problem is ill-defined and its solution requires some additional assumptions. Existing approaches to this problem are generally based on learning, using Factorial Vector Quantization [3], Gaussian (Scaled) Mixture Model [4], or a Hidden Markov Model [5]. In the approach considered here, the sound is modeled as a sum of Auto-Regressive (AR) processes with an additive white noise. Each source is assumed to have a quasi-periodic nature wich makes its parameters identifiable. By quasi-periodic we mean that the source is not exactly periodic but the signal in consecutive periods is almost the same. The easiest way to model such small variations is with a sochastic signal model, the simplest one being a (zero mean) Gaussian Long-Term (LT) AR process. We furthermore superpose a Short-Term (ST) AR aspect to capture the spectral envelope [6, 7]. Such ST+LT AR source models are frequently used in speech encoding algorithms like CELP and LPC [8]. The LT AR part allows to model the quasi-periodic nature of the source. The ST AR model allows to model its spectral envelope, which renders strongly overlapping harmonics of different sources separable.

In [9] we have presented a separation algorithm for such signal model in which the sources and their parameters are estimated jointly via an EM style approach. Due to the Gaussian model, the source extraction consists of Wiener filtering, which gives good results when the model parameters are well estimated. In the EM approach, the AR source parameters are estimated by linear prediction on the reconstructed source correlations, which are the sum of the sample correlations of the estimated source plus the correlation of the source estimation error (orthogonality property of LMMSE estimation). In this paper, we focus on direct estimation of the source parameters from the mixture (sample correlation or sample spectrum), assuming again the number of sources known.
This paper is organized as follow. In section 2 we present the signal model. In section 3 we present the method to estimate the parameters for all sources jointly, while in section 4 we provide a per source approach. Then, in section 5, we provide some simulation results.

2. MODEL

2.1. Signal Model

The model for the sum y_t of short plus long-term autoregressive (AR) Gaussian sources $x_{k,t}$ plus Gaussian white noise v_t (all independent) is :

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

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

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

where t denotes discrete time, K is the number of sources $x_{k,t}$, v_t has variance σ_v^2, $e_{k,t}$ is the excitation signal of source k and is also assumed to be white 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 or even by rounding to the nearest sample if the sampling frequency is high enough), b_k its long-term prediction coefficient and the short-term prediction, of order p_k, uses coefficients $a_{k,n}$.

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

EURECOM's research is partially supported by its industrial partners: BMW Group, Cisco Systems, Monaco Telecom, Orange, SAP, SFR, STEricsson, Swisscom, Symantec, Thales. This research has also been partially supported by the SELIA project of the Futur & Ruptures 08 program of Institut Télécom.

$$A_k(f) = \sum_{n=0}^{p_k} a_{k,n}\, e^{-j2\pi f n} \quad (4)$$

$$B_k(f) = 1 - b_k\, e^{-j2\pi f \tau_k} \quad (5)$$

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

$$S_k(f;\theta_k) = \frac{\sigma_k^2}{|A_k(f)\, B_k(f)|^2}\ ,\ \ k = 1, \ldots, K \quad (6)$$

$$S_0(f;\theta_0) = \sigma_v^2 = \sigma_0^2 \quad (7)$$

The additive noise can be considered as a (short-term) AR model of order 0 and included in the signal set (extension to a more general AR model is immediate).

The source separation algorithm is based on the assumption that the sources can be extracted from the mixture using the knowledge of the parameters, which requires good estimates of these parameters.

2.2. Parameter Subsets

If the parameters can be considered constant during a short time segment we can use a frame based method (length N). The short and long-term aspects of the signals being different by nature, it may seem natural to separate their estimation. The parameters being source-related, we group them by source; this naturally leads to alternating the parameter estimation by source. The overall set of parameters contains the following subsets (short term and long term parameters):

$$\theta = [\theta_1^T \cdots \theta_k^T]^T,\ \theta_k = [\,\mathbf{a}_k\ \varphi_k]^T \quad (8)$$

$$\mathbf{a}_k = [a_{k,1} \cdots a_{k,p_k}]^T,\ \varphi_{\mathbf{k}} = [\,b_k\ \tau_k\ \sigma_k^2\,]^T. \quad (9)$$

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.

3. JOINT ITAKURA-SAITO DISTANCE MINIMIZATION

Many approaches can be used for estimating the AR coefficients from a mixture. Here we propose to minimize the Itakura-Saito (IS) distance, which allows joint spectrum estimation and approximation. The operations being in the spectral domain, low complexity implementations are possible.

3.1. Itakura-Saito Distance

Consider the IS distance between the periodogram $Y(f) = \frac{1}{N}|y(f)|^2$ and the parametric spectrum $S(f;\theta)$

$$IS = \int df \left[\frac{Y(f)}{S(f;\theta)} - \ln\left(\frac{Y(f)}{S(f;\theta)}\right) - 1\right] \quad (10)$$

where $S(f;\theta) = \sum_{k=0}^{K} S_k(f;\theta_k) = \sum_k \sigma_k^2 S_k'(f;\theta_k)$ with $S_k(f;\theta_k)$ the parametric spectrum of the source k, defined in (6) ($S_0'(f;\theta_0) = 1$). If we consider the gradient of IS with respect to (w.r.t.) parameter θ_i, we obtain:

$$\frac{\partial}{\partial\theta_i}\int df \left[\frac{Y(f)}{S(f;\theta)} - \ln\left(\frac{Y(f)}{S(f;\theta)}\right) - 1\right] = \int df \frac{1}{S(f;\theta)^2}[S(f;\theta) - Y(f)]\frac{\partial S_i(f;\theta_i)}{\partial\theta_i} \quad (11)$$

3.2. Weighted Spectrum Matching

It turns out that the IS gradient is the same as that of Optimally Weighted Spectrum Matching. Indeed, at high window length N, the periodogram $Y(f)$ has as mean the spectrum $S(f;\theta)$ and as variance $S(f;\theta)^2$ (with true θ). Hence the optimally weighted spectrum matching criterion becomes

$$\int df\ \frac{1}{S(f;\theta)^2}\ [Y(f) - S(f;\theta)]^2 \quad (12)$$

Taking the gradient w.r.t. a parameter θ_i in the parametric spectrum $S(f;\theta)$ (and ignoring the dependence of the weighting $\frac{1}{S(f;\theta)^2}$ on θ_i) leads to the same gradient as for the IS distance. The weighting involves the true spectrum $S(f;\theta)$, but can asymptotically be replaced by a consistent spectrum estimator such as appropriate versions of the averaged or smoothed periodogram. In our simulations we just use the periodogram itself.

3.3. Gaussian Maximum Likelihood

For sufficiently long window length, Maximum Likelihood (ML) can be expressed in the frequency domain and the negative Gaussian log likelihood of $y(f)$, which has zero mean and variance $N\, S(f;\theta)$, becomes

$$\int df \left[\frac{Y(f)}{S(f;\theta)} + ln\,(S(f;\theta))\right] \quad (13)$$

which obviously will again give the same gradient as the IS distance. This connection with Gaussian ML provides the right angle of attack for introducing a window in the data.

3.4. Short-term AR Parameters Estimation

We provide here the detailed derivations for the case of a short-term AR model only ($\theta_k = \mathbf{a}_k$). We obtain for source k:

$$\frac{\partial S_k(f;\theta_k)}{\partial A_k^*} = -\,S_k(f;\theta_k)\frac{A_k(f)}{|A_k(f)|^2}\ . \quad (14)$$

This leads to a Yule-Walker like equation with a non zero Right Hand Side (RHS), which needs to be solved iteratively:

$$T\left(r_{k,(0,\cdots,p_k-1)}\right)\ \mathbf{a}_k\ =\ g_{k,(1,\cdots,p_k)} - r_{k,(1,\cdots,p_k)} \quad (15)$$

where T is a symmetric Toeplitz matrix, here filled with the first p_k elements of r_k, $\mathbf{a}_k$ are the short term AR coefficients (9), r_k and g_k are defined by:

$$r_k = F^{-1}\left(\frac{Y(f)}{S(f;\theta)}\frac{S_k(f;\theta_k)}{S(f;\theta)}\frac{1}{|A_k|^2}\right) \quad (16)$$

$$g_k = F^{-1}\left(\frac{S_k(f;\theta_k)}{S(f;\theta)}\frac{1}{A_k^*}\right) \quad (17)$$

where F is the Discrete Fourier Transform matrix (we approximate the frequency domain integrals by DFT domain sums).

3.5. Source Power Estimation

Consider equivalently the weighted least squares spectrum matching, weighted by the inverse squared periodogram. We obtain:

$$\int df \frac{1}{Y(f)^2}\left[\sum_{k=0}^{K}\sigma_k^2 S_k'(f;\theta_k) - Y(f)\right]^2\ . \quad (18)$$

The minimization with respect to $\underline{\sigma}^2 = [\ \sigma_v^2\ \sigma_1^2 ... \sigma_K^2\ ]^T$ leads to solving the system $G\ \underline{\sigma}^2 = d$, with:

$$G_{ik} = \int df\ \frac{S_i'(f;\theta_i)\ S_k'(f;\theta_k)}{Y(f)^2}\ ,\ d_i = \int df\ \frac{S_i'(f;\theta_i)}{Y(f)} \qquad (19)$$

3.6. Overall iterative process

Since the r_k and g_k are both A_k dependent, an iterative algorithm is required, which can be summarized as:

- For all the sources k,
- construct r_k and g_k using (16) and (17);
- estimate $\mathbf{a}_k$ by solving (15), construct $S_k(f;\theta_k)$ and g_k using (6) and (17), update $S(f;\theta)$ and $\underline{\sigma}^2$;
- stop condition on r_k, g_k.

The procedure is stopped if the variation between two consecutive estimated correlations is lower than a threshold or if the number of iterations is greater than a maximum number.

4. PER SOURCE WEIGHTED ITAKURA-SAITO DISTANCE MINIMIZATION

In order to find good initial estimates for the joint approach, we shall consider the minimization of a weighted Itakura-Saito distance for the spectrum of one source k, in which the weighting $C_k(f)$ focuses on the harmonics where the source spectrum is much stronger than that of the rest of the signal. The weighted Itakura Saito distance for source k is:

$$\int df\ C_k(f) \left[\frac{Y_w(f)}{S_k(f;\theta_k)} - 1 - ln\left(\frac{Y_w(f)}{S_k(f;\theta_k)} \right) \right] . \qquad (20)$$

At this point we acknowledge the effect of a window in the frame processing. The effect of a window on the spectrum of a short+long term AR model is roughly equivalent to the effect of the long-term correlation coefficient b. Hence, when the long-term correlation is mainly limited by the window, we shall take b arbitrarily close to 1, but incorporate the effect of the window on the spectrum. The source spectrum becomes a sum of harmonic peaks with a fundamental frequency f_0, convolved with the squared Fourier transform $W(f)$ of the (properly normalized) analysis window w_t :

$$\hat{S}_k(f) \quad = \quad \sum_n \alpha_n\, W(f\ -\ n\ f_0) \qquad (21)$$

where the summation range can go up to $\lfloor \frac{0.5}{f_0} \rfloor$ (where f_0 is assumed to be expressed relative to the sampling frequency) or this initial spectral analysis may be limited to a limited frequency range. The spectral peak magnitudes α_n can be seen to be the samples (at frequencies $n\ f_0$) of the spectral envelope which can be modeled as (short-term) AR. The α_n can be estimated by a least-squares fit between $\hat{S}_k(f)$ and $Y_w(f) = \frac{1}{N}|y_w(f)|^2$, the periodogram of the windowed signal $w_t y_t$. The spectrum of the other signals in the mixture can be obtained as the residual spectrum $E_k(f) = \max(Y_w(f) - \hat{S}_k(f;\theta_k), \hat{\sigma}_n^2)$. To improve the spectral estimate w.r.t. a simple residual, we floor the residual at the noise level. The (white) noise level can be estimated from the sorted periodogram values $Y_w(f)$ (after some experimenting, we have taken the value at 20% from the minimum).

4.1. Pitch Estimation

The estimation of the α_n by a least-squares fit between $\hat{S}_k(f)$ and $Y_w(f)$ mentioned above leads to the α_n estimates as simple (scaled) samples of $\hat{S}_k(f) * \breve{W}(f)$ (convolution) at $f = n\ f_0$. The fundamental frequency estimate is then obtained from

$$\hat{f}_{0,k} \quad = \quad \arg\max_f \int df\ \frac{\hat{S}_k(f)}{E_k(f)}\ . \qquad (22)$$

In other words, only the spectral peaks of a source that are less perturbed by the rest of the signal mixture are accounted for. The pitch estimation requires an exhaustive search over the useful frequency range. It can be carried out on a limited range of the spectrum. Multiple pitches are obtained if the cost function (22) shows multiple maxima.

4.2. AR coefficients estimation

An estimate of the short term AR spectral envelope model of source k can be obtained from (20) using the following weighting function:

$$C_k(f) \quad = \quad \frac{Y_w(f)}{E_k(f)}\ . \qquad (23)$$

This weighting focuses the IS distance on frequencies where a single source model is valid. It is assumed though that the resulting subset of frequencies is sufficient to determine the AR spectral envelope correctly, although the estimation quality of the short-term parameters is less critical than that of the long-term parameters (mainly pitch). Minimizing the weighted IS distance leads to an algorithm similar to the one presented in section 3.6 but now both g_k and r_k involve the weighting function.

In the case of an appropriately chosen window (see [9]), the windowing can be expected to dominate the long-term correlation, leading to the following modification of the short-long term AR model

$$S_k(f) = \frac{\sigma_k^2}{|A_k(f)|^2 |B_k(f)|^2} \rightarrow S_k(f) = \frac{\sigma_k^2}{|A_k(f)|^2} \sum_n W(f - n f_{0,k}). \qquad (24)$$

So in this case the source parameters are limited to $f_{0,k}$, $\mathbf{a}_k$ and σ_k^2.

5. SIMULATIONS

5.1. Synthetic data

The first simulation consists of applying the algorithm of section 3.6 on a synthetic spectrum, defined as $Y(f) = \sum_k S_k(f;\theta_k) + \sigma_n^2$ in which we know perfectly the long term parameters ($f_{0,k}$ and b_k). The Short Term AR coefficients and the variances are initialized using the per source approach. The spectral shapes have been chosen to have the same formant frequencies, which is the most difficult case. The result is shown in Fig 1. As the signal is synthetic, and corresponds to the model, the result is almost perfect.

5.2. Real Speech Segment

The next simulation involves a frame of two english speakers, a man and a woman. The lenght of the segment is $64\ ms$ at $8\ KHz$, the (equal power) mixture is artificially made and the signal to noise ratio (SNR) is fixed to $20\ dB$, the periods are estimated using the per source pitch estimator. We use the ensuing parameter estimates for performing the separation (as in [9]), we compare the obtained sources to the original sources and the sources extracted using the

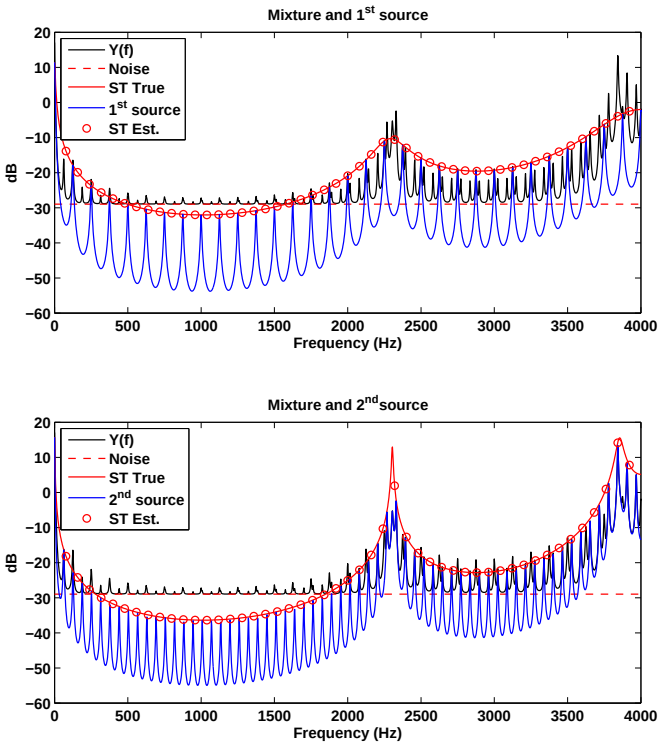

Fig. 1. Spectrum of a synthetic mixture, noise spectrum, true source spectra, true source spectral envelopes, estimated sources spectral envelopes.

parameters estimated on the individual source signals (before the mixing process). The separation algorithm from [9] extracts windowed source frames. The waveform of the decomposition is shown in Fig 2. The difference between the two source extraction versions is small. Note that the speech segments are well voiced.

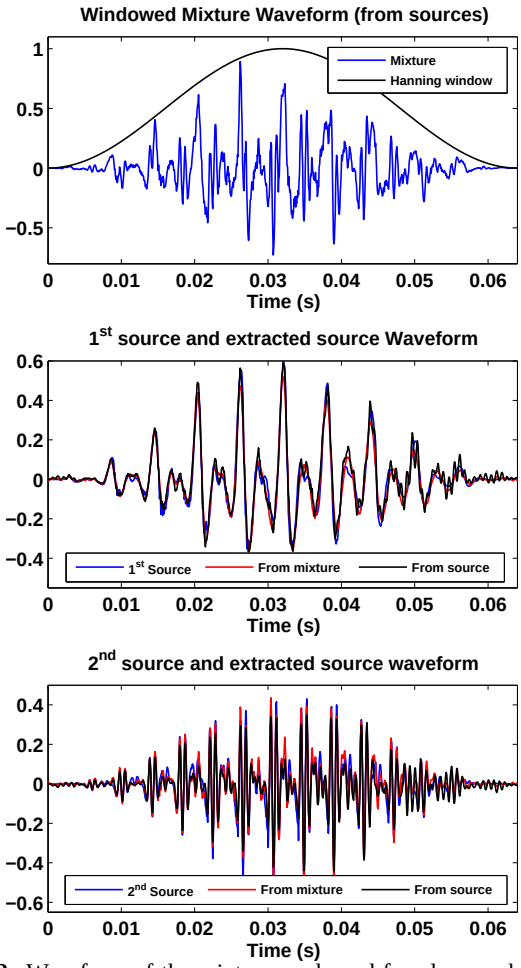

Fig. 2. Waveform of the mixture, male and female speech sources and estimated sources. The sources are extracted with the parameters determined from the mixture or from the source.

6. CONCLUSIONS

In this paper we have proposed an algorithm, based on the minimization of the Itakura-Saito (IS) distance, for estimating the short+long term AR parameters of several sources and also the additive noise variance from a mixture. The minimization of IS leads to an iterative algorithm involving Yule-Walker like equations for the AR aspect and weighted least-squares spectrum matching for the estimation of the powers of the sources. We have also presented an algorithm based on the Weighted Itakura-Saito distance for the initizalition of the parameters in which we provide algorithms for source pitch and AR spectral envelope estimation on a windowed version of the data. Simulations on synthetic and real data are very encouraging. The estimated parameters lead to separation results that are close to the separation obtained by using the parameters estimated on the individual sources. Future work will include the integration of the window in the joint algorithm.

7. REFERENCES

[1] A. Hyvarinen, "Survey on independent component analysis," 1999, Neural Computing Surveys.

[2] P. Comon and C. Jutten, "Handbook of blind source separation: Independent component analysis and applications," 2010, Academic Press.

[3] S.T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*, 2000.

[4] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. on Audio, Speech Language Processing*, 2007.

[5] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *ICA03*, Nara, Japan.

[6] A Schutz and D.T.M Slock, "Blind audio source separation using short+long term AR source models and iterative itakura-saito distance minimization," in *IWAENC*, 2010.

[7] S Bensaid, A Schutz, and D.T.M Slock, "Single Microphone Blind Audio Source Separation Using EM-Kalman Filter and Short+Long Term AR Modeling," in *LVA*, 2010.

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

[9] A Schutz and D.T.M Slock, "Single-microphone blind audio source separation via Gaussian Short+Long Term AR Models," in *ISCCSP*, 2010.