SINGLE SENSOR SOURCE SEPARATION USING MULTIPLE-WINDOW STFT REPRESENTATION

[1]*Laurent BENAROYA,* [1]*Raphaël BLOUET,* [1]*Cédric FÉVOTTE and* [2]*Israel COHEN*

[1] MIST Technologies Research
204, rue de Crimée
75019 Paris, France
firstname.lastname@mist-technologies.com

[2] Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
icohen@ee.technion.ac.il

ABSTRACT

The aim of this paper is to investigate the use of multiple-window Short-Time Fourier Transform (STFT) representation for single sensor source separation. We propose to iteratively split the observed signal into target sources and residuals. Each target source is modeled as the sum of elementary components with known power spectral densities (PSDs). The approach involves a non negative decomposition of the spectra of the observed mixure in a given frame into a dictionary of PSDs. The resolution of the PSDs varies at each iteration of the algorithm. The decomposition into source signals and residual signals employs a confidence measure, which is based on the Fisher information matrix of the expansion coefficients. We demonstrate the improved performance of the proposed method on mixtures of voice and music signals.

1. INTRODUCTION

Recently, we introduced a single sensor blind source separation approach [1], which is based on an extension of the Wiener filtering to nonstationary processes through the use of Gaussian mixture models. The analysis is carried out in the joint time-frequency domain with the Short-Time Fourier Transform (STFT) of the signals. In this domain, we have defined the notion of an elementary source $\mathcal{S}s_k(t,f) = \sqrt{a_k(t)} \cdot \mathcal{S}b_k(t,f)$, where $\mathcal{S}$ is the STFT operator, $a_k(t)$ is a non negative amplitude parameter and b_k is a zero-mean stationary Gaussian process with diagonal covariance matrix $\Sigma_k = \{\sigma_k^2(f)\}_f$. The amplitude parameter can be seen as either a temporal envelope parameter or an activation parameter. We define a composite source as the sum of independent elementary sources over a set of indices $K_i : \mathcal{S}s_i(t,f) = \sum_{k \in K_i} \sqrt{a_k(t)} \cdot \mathcal{S}b_k(t,f)$.

This work has been partially supported by the European Commission's IST program under project Memories.

The resulting separation algorithm, in a Bayesian framework, consists of two steps :

1. Computation of the amplitude parameters $\{a_k(t)\}$ in a maximum likelihood sense, for all frame indices t.
2. Filtering the original mixture according to the resulting adaptive Wiener filters.

In this paper, we propose an improvement of this algorithm by using a multiple-window STFT representation. The basic idea is to decompose, in an iterative fashion, the observed signal into source components and residual components for several window lengths. The algorithm initializes with a relatively long window, which becomes shorter throughout the iterations. The source components in each iteration are defined as the components that are well represented in the current resolution. The input signal for iteration i is the residual generated at the iteration $i-1$.
The paper is organized as follows. In Section 2, we briefly recall the general framework of the mono-resolution algorithm presented in [1]. In Section 3, we introduce the new approach with a special emphasis on the choice of a confidence measure. This measure is used for selecting the subset K_i that corresponds to the most reliable amplitude parameters. In Section 4, we present experimental results obtained on mixtures of music and voice signals. Conclusions and perspectives are given in Section 5.

2. OVERVIEW OF THE CLASSICAL ALGORITHM

2.1. Notations

Let x denote the observed signal, which is assumed to be a mixture of two unknown source signals, s_1, s_2. Let $\mathcal{S}$ denote the STFT operator and $\mathcal{S}x(t,f)$ the STFT of $x(n)$, where n is a discrete time index, t and f are respectively the time-frame and frequency-bin indices. Then, in

the STFT domain we have

$$\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f)\,.$$

Note that we restrict ourselves here to two sources although the generalization to more than two sources is theoretically straightforward and has been successfully tested.

2.2. Learning the PSDs sets

We assume that we have some clean training samples of each source. These training excerpts do not need to include the source signals contained in the observed mixture, but we assume that they sufficiently diverse and characterize the type of expected source signals. For example we may learn elementary drums PSDs on a range of drums solos. From these training samples, we estimate the covariance matrices (or PSDs set) $\{\sigma_k^2(f)\}_{k\in K_i}$ for each source s_i. We may use, for instance, a vector quantization algorithm on the short-time Fourier spectra of the excerpts in order to build the PSDs set.

2.3. Amplitude parameters estimation

Conditionally upon the amplitude parameters $\{a_k(t)\}_k$, all the elementary sources $\mathcal{S}s_k(t,f)$ are independent zero-mean Gaussian processes with variance $\{a_k(t)\sigma_k^2(f)\}$. Then the observed mixture is also a zero-mean Gaussian process with variance $\{\sum_k a_k(t)\sigma_k^2(f)\}$. Therefore, we have the following log-likelihood equation :

$$\log p(\mathcal{S}x(t,f)|\{a_k(t)\}) = \\ -\frac{1}{2}\sum_f \left[\frac{|\mathcal{S}x(t,f)|^2}{E(t,f)} + \log(E(f,t))\right],$$

where $E(t,f) = \sum_{k\in K_1\cup K_2} a_k(t)\sigma_k^2(f)$. We can estimate the amplitude parameters $\{a_k(t)\}_k$ by setting the first derivative of the log-likelihood to zero under a non negativity contraint. As this problem has no analytic solution, we use an iterative fixed-point algorithm with multiplicative updates [2, 3, 4], yielding

$$a_k^{(\ell+1)}(t) = a_k^{(\ell)}(t)\cdot\frac{\sum_f \sigma_k^2(f)\cdot\frac{|\mathcal{S}x(t,f)|^2}{E^{(\ell)}(t,f)^2}}{\sum_f \sigma_k^2(f)\cdot\frac{1}{E^{(\ell)}(t,f)}},$$

where $E^{(\ell)}(t,f) = \sum_k a_k^{(\ell)}(t)\sigma_k^2(f)$.

2.4. Source separation

Conditionally upon the estimated amplitude parameters $\{a_k(t)\}_k$, sources separation can be obtained through a generalized Wiener formula :

$$\widehat{\mathcal{S}s_i}(t,f) = \frac{\sum_{k\in K_i} a_k(t)\sigma_k^2(f)}{\sum_{k\in K_1\cup K_2} a_k(t)\sigma_k^2(f)}\mathcal{S}x(t,f)\,.$$

3. THE MULTIRESOLUTION APPROACH

3.1. General description of the algorithm

Let $w_1(n),\ldots,w_N(n)$ denote N windows with decreasing support length, and let $\mathcal{S}_{w_i}$ denote the STFT operator with an analysis window $w_i(n)$. We first basically apply the algorithm of Section 2 with the longest window $w_1(n)$. This algorithm is slightly modified to yield a residual signal r_{w_1}, such that

$$\begin{aligned}\mathcal{S}_{w_1}x(t,f) &= \mathcal{S}_{w_1}s_{1,w_1}(t,f) + \mathcal{S}_{w_1}s_{2,w_1}(t,f) \\ &\quad + \mathcal{S}_{w_1}r_{w_1}(t,f)\,.\end{aligned}$$

After inverse STFT, we iterate on $r_{w_1}(n)$ with analysis window w_2, and so on. At the iteration i we have

$$\begin{aligned}\mathcal{S}_{w_i}r_{w_{i-1}}(t,f) &= \mathcal{S}_{w_i}s_{1,w_i}(t,f) + \mathcal{S}_{w_i}s_{2,w_i}(t,f) \\ &\quad + \mathcal{S}_{w_i}r_{w_i}(t,f)\,.\end{aligned}$$

While no residual signal is computed with the monoresolution approach, the multiresolution approach involves the selection of a set of PSDs with their associated amplitude parameters. This is done through a partition of the amplitude parameters indices $k\in K_1\cup K_2$ into three different sets $Q_1(t)$, $Q_2(t)$ and $R(t)$. The set $R(t)$ contains the indices k such that the corresponding $\{a_k(t)\}_{k\in R(t)}$ are "unreliably" estimated and the set $Q_1(t)$ (resp. $Q_2(t)$) contains the indices $k\in K_1$ (rep. $k\in K_2$) of reliably estimated $a_k(t)$. More precisely, this partition is carried out through the computation of a confidence measure $J_k(t)$. This confidence measure should be high if the corresponding estimate of $a_k(t)$ is reliable. As will be seen in Section 3.2, the confidence measure that we have chosen is related to the Fisher information matrix of the likelihood of the amplitude parameters.

Note that these three sets of indices $Q_1(t)$, $Q_2(t)$ and $R(t)$ are frame dependent. Relying on similar filtering formulae as those used in the classical algorithm, we obtain three estimates $\hat{s}_{1,w_i}(n)$, $\hat{s}_{2,w_i}(n)$ and $\hat{r}_{w_i}(n)$ (back in the time domain). Then we can iterate on $\hat{r}_{w_i}(n)$ with a different STFT window $w_{i+1}(n)$.

Finally, we have the following estimates,

$$\hat{s}_1(t) = \sum_{i=1}^{N}\hat{s}_{1,w_i}(t)$$

$$\hat{s}_2(t) = \sum_{i=1}^{N} \hat{s}_{2,w_i}(t)$$
$$\hat{r}(t) = \hat{r}_{w_N}(t).$$

With this algorithm we expect that long signal components will be reliably estimated with long analysis windows, while short signal components, such as transients, will remain in the residual signal until the window length is sufficiently small to capture them reliably.

3.2. Choice of a confidence measure

Suppose we have a confidence interval on each amplitude parameter $a_k(t) : a_k(t) \in [\hat{a}_k(t) - \ell_k(t); \hat{a}_k(t) + L_k(t)]$. Then the quantity $J_k(t) = \hat{a}_k(t)/\left[L_k(t) - \ell_k(t)\right]$ can be considered as a relative confidence measure for the estimate $\hat{a}_k(t)$. If $J_k(t) > \lambda$ where λ is an experimentally tuned threshold, we assume that the estimate $\hat{a}_k(t)$ at frame t is reliable. Using a Taylor expansion of the log-likelihood around the ML estimate, we have

$$\log p(r_{w_i}|\{\hat{a}_k(t) + \delta a_k(t)\}_k) \approx \log p(r_{w_i}|\{\hat{a}_k(t)\}_k) - \frac{1}{2}[\delta a_k(t)]^T H(t)[\delta a_k(t)],$$

where $H_{i,j}(t) = -\frac{\partial^2}{\partial a_i(t)\partial a_j(t)} \log p(r_{w_i}|\{\hat{a}_k(t)\}_k)$. Taking the expectation on both sides of the approximate equality, we get

$$E\left(\log \frac{p(r_{w_i}|\{a_k(t)\})}{p(r_{w_i}|\{a_k(t) + \delta a_k(t)\})}|\{a_k(t)\}\right) \approx \frac{1}{2}[\delta a_k(t)]^T I(t)[\delta a_k(t)],$$

where the left side of the approximate equality is simply the Kullback-Leibler divergence and $I(t)$ is the Fisher information matrix for $a_k(t) = \hat{a}_k(t)$. This relationship is well known and is also true if $\{a_k(t)\}_k$ is not a local optimum [5]. For a given admissible error E on the Kullback-Leibler divergence, we get

$$|\delta a_k(t)| \leq \sqrt{2E} \cdot \sqrt{[I^{-1}(t)]_{k,k}},$$

thus yielding a confidence interval on $a_k(t)$ for a given admissible error E on the objective function.

Note that the *sensitivity* of the estimated parameters to a small change in the objective function (here, the log-likelihood) or a mis-specification of the objective function is related to the inverse of the Fisher information matrix. In our model, the Fisher information matrix is $I_{i,j}(t) = \frac{1}{2}\sum_f \frac{\sigma_i^2(f)\sigma_j^2(f)}{E(t,f)^2}$. Take the inverse of $I(t)$, we obtain

$$J_k(t) = \frac{\hat{a}_k(t)}{\sqrt{[I^{-1}(t)]_{k,k}}}.$$

3.3. Practical choice of the thresholds

As mentioned above, it is possible to tune the thresholds in an experimental way. A way to circumvent this problem is to sort the confidence measures $J_k(t)$ for each fixed frame index t. Then we can keep the M more reliable estimates for each frame, and use all the other k estimates to build the residual. An alternative way to proceed is to build the residual set $R(t)$ by taking the less reliable indices such that $\sum_{k\in R(t)} a_k(t) < \epsilon \sum_{k\in K_{1,w_i}\cup K_{2,w_i}} a_k(t)$, for a given $\epsilon \in [0, 1]$. Indeed, the first sum is the estimated variance of the residual r_{w_i} while the second sum is the estimated variance of the overall decomposition $r_{w_{i-1}}$ for each frame. Then after N iterations, the residual variance is approximately lower than ϵ^N times the original signal variance.

4. EXPERIMENTAL STUDY

4.1. Experimental protocol

The evaluation task consists of unmixing a voice plus jazz music audio track. All the audio excerpts are sampled at 16 kHz. We make a 15 seconds long linear mix of a male voice in French and an excerpt of a jazz piece with 0 dB Signal to Noise Ratio (SNR). The voice excerpt has been recorded under good environmental conditions. The voice PSDs are trained on a set of about 50 short excerpts of various male speakers. The jazz piece is an excerpt of *The Four Seasons* by the *Jacques Loussier Trio*. The excerpt contains piano, bass and drums. We were given training data for each instrument. Using a Vector Quantization (VQ) algorithm, the training step had to be done for each window size, namely 64, 16 and 8 ms. We obtain respectively :

- 83, 113 and 180 PSDs for the piano,
- 56, 81 and 89 PSDs for the bass,
- 9, 30 and 59 PSDs for the drums,
- 289, 369 and 453 PSDs for the speech model.

4.2. Evaluation criteria

The criteria we use for the separation performance are described in [6]. Basically, the SDR (Source to Distortion Ratio) provides an overall separation performance criterion,

the SIR (Source to Interference Ratio) measures the level of the interferences from other sources in each source estimate and the SAR (Signal to Artifacts Ratio) measures the level of artifacts in the source estimates. The higher are the ratios, the better is the quality of the estimation.

4.3. Evaluation

In this evaluation, we present the SDR, SIR and SAR results on three different configurations, for both instrumental and speech parts. The first configuration is the standard pseudo-Wiener algorithm with a single STFT window of length 16 ms. The second configuration uses the modified algorithm with two STFT windows of size 64 and 8 ms. Finally, the third configuration uses three windows of length 64, 16 and 8 ms.[1]

TAB. 1 – SDR, SIR and SAR for the different methods using a 0 dB SNR mixture.

source	SDR	SIR	SAR
1 STFT window			
music	3.8	6.1	8.6
voice	-1.9	5.1	0.1
2 STFT windows			
music	4.2	6.8	8.6
voice	-2.3	10.1	-1.7
3 STFT windows			
music	4.2	7.4	7.7
voice	-2.3	9.7	-1.0

As can be seen in Table 1, the SDR is slightly improved with the new method on the music part (around 0.4 dB) but the improvement is not clear in the speech component case. Moreover, the figures are very similar with two and three windows. The small improvement in the SDR for the music component when we use more than one STFT window is confirmed by the SIR and SAR scores. The case of the speech is different. Indeed, we have an improvement of 5 dB in SIR from one window to two windows. This improved SIR is obtained at the cost of a lower SAR. However, we have listened to the separated speech components with the one and two STFT windows methods, and we have noticed that the intelligibility of the speech is improved with the new method, although the SAR decreases.

[1] Audio samples are available at `http://persos.mist-technologies.com/~lbenaroy/iwaenc/`.

4.4. Discussion

In order to understand the practical problem we had to deal with, it should be noticed that in many cases, the local energy of the mixture is spread over just a few amplitude parameters (usually no more than 4 $a_k(t)$ per frame, among several hundreds of them). Therefore splitting the signal in source components and a residual component is a rather difficult task. A way to avoid this phenomenon, would be to add, in a future work, a prior density on each amplitude parameter.

5. CONCLUSION

We have proposed a single sensor audio source separation method, which is based on multiple-window STFT representation and a recently introduced audio source separation method. The proposed approach facilitates analysis of sound events at different time scales and thus leads to enhanced separation performance. A few observations worth to be mentioned. First, we could replace the iterative STFT scheme with a hierarchical multi-window analysis, as is done with local cosine packet. Second, we can use prior densities on the amplitude parameters in order to estimate the residual signal more easily. Finally, this algorithm could be used to get a multiresolution segmentation of a composite audio signal.

6. REFERENCES

[1] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, January 2006.

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advanced Neural Information Proceecing Systems*, 2001, vol. 13, pp. 556–562.

[3] P. O. Hoyer, "Non-negative sparse coding," in *Proc. IEEE Workshop on Neural Networks for Signal Processing, Martigny, Switzerland*, 2002, pp. 557–565.

[4] L. Benaroya, *Séparation de plusieurs sources sonores avec un seul microphone*, Ph.D. thesis, Université Rennes 1, 2003.

[5] C. Arndt, *Information Measures*, Springer, 2001.

[6] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *ICA*, Nara, Japan, 2003, pp. 715–720.