SINGLE SENSOR SOURCE SEPARATION BASED ON WIENER FILTERING AND MULTIPLE WINDOW STFT

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

[1] MIST Technologies Research Team
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 multiresolution framework for single sensor source separation based on pseudo-Wiener filtering. We propose a scheme in which the signal is iteratively split in target sources and a residual. Each target source is modeled as the sum of elementary components with known Power Spectral Densities (PSDs). The approach boils down to perform a non negative decomposition of the spectra of the observed signal in a given frame onto the dictionnary of known PSDs. The resolution of the PSDs (and hence the frame length) is changed at each iteration of the algorithm. The decomposition into sources plus residual is done thanks to a confidence measure based on the Fisher information matrix of the expansion coefficients. After theoretical developments we compare the mono and multiresolution approaches and a set of audio examples.

1. INTRODUCTION

In [1] we proposed a generalization of the well-known Wiener filter [2] to locally stationary audio sources. The analysis is done in the time-frequency plane through the Short Term 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 $\mathcal{S}b_k(\cdot,f)$ is a zero mean, stationary, Gaussian process with diagonal covariance matrix $\Sigma_k = \{\sigma_k^2(f)\}_f$. The amplitude parameter can either be seen as a temporal envelope parameter or a activation parameter. We define a composite source as the sum of independent elementary sources over a set of indices K_i : $\mathcal{S}s_i(t,f) = \sum_{k\in K_i} \sqrt{a_k(t)} \cdot \mathcal{S}b_k(t,f)$.
The resulting separation algorithm (in a Bayesian framework) consists of two steps :

1. Compute the amplitude parameters $\{a_k(t)\}$ in a Maximum Likelihood sense, for all frame indices t.
2. Filter the original mixture according to the resulting adaptive Wiener filters.

In this paper, we propose an improvement of this algorithm using a multiresolution STFT scheme. The basic idea is to decompose, in an iterative fashion, the observed signal into source components and a residual for several window lengths. At each iteration the residual contains components that are not properly represented at the current resolution. The input signal at iteration i is then the residual generated at iteration $i-1$. The algorithm starts with a long window sizes wich is decreased throughout the iterations.
The paper is organized as follows. In Section 2 we recall the general framework of the monoresolution algorithm presented in [3]. In Section 3 we explain in details the new approach with a special emphasis on the choice of the confidence measure. This measure is used to select a subset of each set K_i that corresponds to accuratly estimated amplitude parameters. In Section 4 we present results obtained on a mixture a music and voice. Conclusions and perspectives are given in section 5.

2. OVERVIEW OF THE CLASSICAL ALGORITHM

2.1. Notations

We note $\mathcal{S}$ the STFT operator and $\mathcal{S}x(t,f)$ is the STFT of $x(n)$, where n is the discrete time domain index, t and f are respectively the discrete frame index and frequency bin. We have the following observation equation

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

where x is the observed mixture and s_1, s_2 are the unknown sources. 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 be identical to the source contained in the observed mixture but we assume that they are "representative" of the source. For example we might 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 term 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 :

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

where $en(f,t) = \sum_{k\in K_1\cup K_2} a_k(t)\sigma_k^2(f)$. 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 [4, 5, 3], yielding

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

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

2.4. Sources estimation

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

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

Note that this estimator is equivalent to the MAP estimator under Gaussian assumptions.

3. THE MULTIRESOLUTION APPROACH

3.1. Notations

We suppose that $w_1(n),\ldots,w_N(n)$ are N windows with decreasing support length. We note $\mathcal{S}_{w_i}$ the STFT operator with analysis window $w_i(n)$.

3.2. General description of the algorithm

We first basically apply the algorithm of Section 2 with the longest window $w_1(n)$. This algorithm is slightly modified as to yield a residual signal, such that

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

After inverse-STFT, we iterate on $r_1(n)$ with analysis window w_2. At the end of the day, the decomposition at iteration i is

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

While no residual 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 $a_k(t)$. At each step, sources and residual are estimated with :

$$\begin{aligned}
\widehat{\mathcal{S}s}_{1,w_i}(t,f) &= \frac{\sum_{k\in Q_1(t)} a_k(t)\sigma_k^2(f)}{en(t,f)}\mathcal{S}r_{w_{i-1}}(t,f)\\
\widehat{\mathcal{S}s}_{2,w_i}(t,f) &= \frac{\sum_{k\in Q_2(t)} a_k(t)\sigma_k^2(f)}{en(t,f)}\mathcal{S}r_{w_{i-1}}(t,f)\\
\widehat{\mathcal{S}r}_{w_i}(t,f) &= \frac{\sum_{k\in R(t)} a_k(t)\sigma_k^2(f)}{en(t,f)}\mathcal{S}r_{w_{i-1}}(t,f)\\
&\quad \text{with } \mathcal{S}r_{w_0}(t,f) = \mathcal{S}x(t,f)
\end{aligned}$$

$Q_1(t)$, $Q_2(t)$ and $R(t)$ are obtained through the computation of a confidence measure $J_k(t)$. This confidence measure should be small if the corresponding estimate of $a_k(t)$ is accurate. As will be seen in Section 3.3, 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 a similar filtering formula than those used in the classical algorithm, we get 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 get the estimates :

$$\begin{aligned}
\hat{s}_1(t) &= \sum_{i=1}^{N}\hat{s}_{1,w_i}(t)\\
\hat{s}_2(t) &= \sum_{i=1}^{N}\hat{s}_{2,w_i}(t)\\
\hat{r}(t) &= \hat{r}_{w_N}(t).
\end{aligned}$$

We this algorithm we expect that short components such as transients will be unreliably estimated with long analysis windows and therefore will fall in the residual untill the window length is sufficiently small to capture them reliably.

3.3. Choice of a confidence measure

Suppose we have a confidence interval on each amplitude parameter $a_k(t) : a_k(t) \in [\hat{a}_k(t) - l_k(t); \hat{a}_k(t) + L_k(t)]$. Then the quantity $J_k(t) = \frac{L_k(t)-l_k(t)}{\hat{a}_k(t)}$ can be seen as a relative confidence measure on the estimate $\hat{a}_k(t)$. If $J_k(t) \leq \lambda$ where λ is a experimentally tuned threshold, we consider that the estimation of $a_k(t)$ at frame index t is reliable. Using a Taylor expansion of the opposite log-likelihood around the ML estimate, we get

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

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

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

where the left side of the equality is simply the Kullback-Leiber 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 [6]. For a given admissible error E on the Kullback-Leiber divergence, we get

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

thus yielding a confidence iterval on $a_k(t)$ for a given admissible error E on the objective function.
Note that we see here that the *sensitivity* of the estimated parameters to a small change of the objective function (here, the opposite 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 $I_{i,j}(t) = \frac{1}{2}\sum_f \frac{\sigma_i^2(f)\sigma_j^2(f)}{en(f,t)^2}$. We have to take the inverse of $I(t)$ for all t and we get

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

3.4. Practical choice of the thresholds

As we said before it is possible to tune the thresholds in an experimental way. A way to circonvent this problem is to sort the confidence measures $J_k(t)$ for each fixed frame index t. Then we can either keep the M more reliable estimates for each frame, all the other k indices beeing used to build the residual. Conversely, another 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, we are garanteed that the residual variance is (approximately) lower that ϵ^N times the original signal variance.

4. EXPERIMENTAL STUDY

4.1. Experimental protocol

The evaluation task consists in unmixing a voice plus jazz music audio track. All the audio excerpts are sampled at 16kHz. We make a 15 seconds long linear mix of a male voice in French and an excerpt of a jazz piece with 0dB Signal to Noise Ratio (SNR). The voice excerpt has been recorded in 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 algorithm (VQ), 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 is described in [7]. 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 section, 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. For both multi-resolution approaches, we select at each step the $M = 5$ most reliable estimates of the amplituds factor coefficients.

SNR = 0 dB			
instrument	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

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

As can be seen on 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 done 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.

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 future work, a prior density on each amplitude parameter. Our approach introduces new parameters (namely M and ϵ) that we need to tune. Futur effort to enhance this approach will consist in automatically set these parameters.

5. CONCLUSION

We have proposed a new single sensor audio source separation method based on a previous pseudo-Wiener method and multiple STFT windows. We believe this contribution is important as it allows to analyse sound events at different time scales and thus to enhance the separation performance. It is a difficult task that is addressed here as we deal with superimposed audio events with different scales. A few perspectives can ben mentionned. First, we could replace the iterative STFT scheme with a hierarchical multi-window analysis, as it 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, R. Gribonval, and F. Bimbot, "Non negative sparse representation for wiener based source separation with a single sensor," in *ICASSP*, Hong Kong, 2003, pp. 613–616.

[2] N. Wiener, *Extrapolation, interpolation and smoothing of stationary time series*, MIT press, 1949.

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

[4] 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.

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

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

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