DOPING AUDIO SIGNALS FOR SOURCE SEPARATION

Gaël Mahé

LIPADE
Université Paris Descartes
Sorbonne Paris Cité
France

gael.mahe@parisdescartes.fr

Everton Z. Nadalin, João-Marcos T. Romano

DSPCom Lab
School of Electrical and Computer Engineering (FEEC)
University of Campinas
Brazil

nadalin@dca.fee.unicamp.br, romano@dmo.fee.unicamp.br

ABSTRACT

This work fits in the frames of sparse component analysis (SCA), informed source separation (ISS) and doping watermarking. The SCA relies on a strong hypothesis of sparsity of the sources. In a particular context where the original sources are available (ISS), we make the distributions of the time-frequency coefficients of the sources more sparse, through a doping watermarking that imperceptibly transform the histogram of the coefficients. Using the "sparsified" sources instead of the original ones in a SCA leads to a better estimation of the number of sources and to a more accurate identification of the mixing system.

***Index Terms*—** audio, sparse component analysis (SCA), informed source separation (ISS), doping watermarking

1. INTRODUCTION

In the case where as many mixtures as sources are available, the Blind Source Separation (BSS) may be performed through the Independant Component Analysis (ICA) [1], relying on the sole hypothesis that the sources are mutually independent. In the under-determined case, *i.e.* when the number or mixtures is lower than the number of sources, another assumption, commonly used in audio source separation, is that the sources are sparse: there are some "gaps" of silence in each source signal. When the signals do not overlap (it is said that the sources have disjoint orthogonality), it is possible to perfectly recover all the sources through Sparse Component Analysis (SCA) [1, 2].

In some signals, audio for example, although there is temporal sparsity, it is usually not enough to perform a good separation [1]. That is why most studies deal with the sources in time-frequency domain. Nevertheless, it is not possible to guarantee the assumption of disjoint orthogonality in most of the cases.

To overcome the strength of this sparsity hypothesis, some works developed the concept of "informed source separation" (ISS) [3, 4, 5], in the particular context where the sources are available. The principle of ISS is to embed in the mixture a watermark describing the sources and the mixing process, that can be extracted by the receiver of the mixture to help the separation from an under-determined mixture.

In [3], the time-frequency plane is divided in "molecules" and the watermark is either the energy contribution of each source to each molecule of the mixture, or a coarse description of each molecule of each source. This watermark helps the separation of a linear instantaneous monophonic mixture of 4 or 5 sources. In the stereophonic case, [4] embeds through watermarking the mixture matrix and, for each molecule, the index of the 0, 1 or 2 sources dominating in the molecule. In reception, thanks to this information, each molecule undergoes the separation process as a (over-)determined mixture.

An ISS based on the modeling of the source signals by "latent components" was proposed in [5] for convolutive under-determined mixtures: the time-frequency bins of each source are a time-varying combination of complex centered gaussian variables. The watermark contains the parameters of the model, the mixing filters and the unmixing filters. Using this information, the separation is performed through a generalized Wiener fitering.

These methods reach good performance, even for very under-determined mixtures, but they require a high rate of watermark (20 kbit/s in [4] to *ca.* 100 kbit/s in [3]), which make them unsuitable for compressed audio (actually, the target application is the audio CD). Another drawback is that the embedded information are restricted to a particular mixture, both because the watermarking is on the mixture (this could be changed) and, more deeply, because the watermark needs the knowledge of the mixing matrix.

Another track in watermarking-aided audio-processing was developed in the same period, namely the "doping watermarking" [6, 7, 8]. The principle is to imperceptibly change the properties of an audio signal, in order to enhance a particular processing. This was used to "stationarize" audio signals to enhance acoustic echo cancelation [6], to "gaussianize" signals for non-linear system identification [7], and to low-

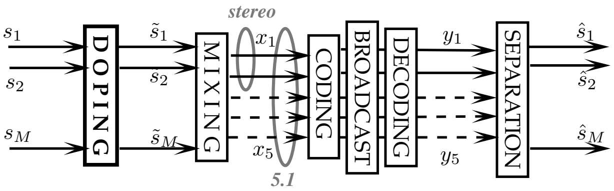


Fig. 1. From doping to separation.

pass filter the probability density function for application of the quantization theorem [8]. Unlike conventional watermarking, the inserted information are not a binary stream, but a particular property. However, the doping watermarking must respect the same inaudibility and robustness constraints.

Coming back to source separation, a suitable doping watermarking could increase the sparsity of the sources, without changing their perceptual characteristics. Hence, any underdetermined mixture of the "sparsified" sources should be easier to separate by SCA techniques. The goal of this paper is to show how the source separation of audio signals can be enhanced by a doping watermarking that imperceptibly "sparsifies" the sources.

The targeted application is the case where audio sources are recorded separately in a studio, mixed in stereo or 5.1 and finally recorded as CD or compressed files for broadcast. As illustrated in Fig 1, each source is doped before the mixing, in order to enhance the SCA from the received mixture. This approach also belongs to ISS, since the separation relies on a preliminary acces to the sources.

A "sparsification" was proposed in [9], which principle is to set to zero the source time-frequency (TF) coefficients under the masking theshold (*ca.* 75 % of the coefficients, without audible distorsion). This was used as a pre-processing step in the ISS described in [4], where an ICA is performed in each TF bin of a stereo mix, based on the assumption of 0 to 2 dominating sources in each bin. It provides a computational gain, since the amount of TF bins with zero source, which do not need to be separated, increases. But, as the authors say, it leads to only few improvement in separation quality, because the bins for which a perfect separation is allowed (0 to 2 sources) represent only 10 % of the energy of the mix.

In the first part, we will propose a another process of "sparsification" and analyze the "sparsified" signals in terms of sparsity and quality. The second part will study how this "sparsification" can help in SCA, principally for identification of the number of sources and of the mixing system.

2. SPARSIFICATION

The distributions of time-frequency coefficients of audio signals may be approximated by complex Generalized Gaussian distributions, with a form factor varying between 0.2 and 0.4 [1]. Our goal is to enhance the sparsity of any audio signal through reducing the form factor of its distribution.

2.1. Principle of sparsification

Denoting β the form factor of the original distribution and $\beta' = \beta/\lambda$, with $\lambda > 1$, the target form factor, the target probability density function of the modulus of the time-frequency coefficients is :

$$f_{target}(|z|) = \frac{\beta'}{\alpha'\Gamma(1/\beta')} \exp\left(- \left|\frac{z}{\alpha'}\right|^{\beta'} \right)$$

where, denoting α the scale factor of the original distribution,

$$\alpha' = \alpha\sqrt{\frac{\Gamma(3/\beta)\Gamma(1/\beta')}{\Gamma(1/\beta)\Gamma(3/\beta')}}$$

in order to maintain the same variance.

The principle of sparsification is similar to histogram equalization in image processing or "gaussianization" in [7]. Sparsifying the audio signal means transforming each time-frequency coefficient modulus $|S(m,k)|$ into $|\tilde{S}(m,k)|$, so that :

$$F_{target}(|\tilde{S}(m,k)|) = F_{emp}(|S(m,k)|)$$

where F_{target} is the cumulative distribution function associated to f_{target} and F_{emp} is the empirical cumulative distribution function computed from $\{|S(m,k)|\}$.

2.2. Algorithm

1. Compute the time frequency representation $S(m,k)$, using non-overlapping windows of length 32 ms;
2. Estimate the form factor β of the distribution of $|S(m,k)|$ through the moments method [10], assuming a Generalized Gaussian distribution;
3. Fix the target form factor $\beta' < \beta$ and $\forall\, m,k$ compute $|\tilde{S}(m,k)|_0 = F_{target}^{-1}\big(F_{emp}(|S(m,k)|)\big)$;
4. Implement in the time domain (see subsection 2.3) the transformation $|S(m,k)| \rightarrow |\tilde{S}(m,k)|_0$, leading to a signal $\tilde{s}$ with time-frequency representation $|\tilde{S}(m,k)|$

2.3. Implementation in the time domain

For each m^{th} frame of length N, the ratio $|\tilde{S}(m,k)|_0/|S(m,k)|$ gives the frequency response $|H(m,k)|$ of the filter that must be applied to the frame. The m^{th} frame of the sparsified signal $\tilde{s}$ is computed as follows

1. Symetrize the impulse response $\text{DFT}^{-1}(|H(m,k)|)$ to get a linear phase filter $h(m,n)$ with maximum value for $n = (N-1)/2$.
2. Compute FFT of $h(m,n)$ on $2N$ samples $\rightarrow H_{2N}(m,k)$

3. Concatenate the m^{th} frame of s with the second half of the preceding and the first half of the next $\rightarrow s'_m$

4. Compute FFT of $s'_m \rightarrow S_{2N}(m,k)$

5. Get $\tilde{S}_{2N}(m,k) = H_{2N}(m,k)S_{2N}(m,k)$

6. The m^{th} frame of $\tilde{s}$ equals the second half of $\text{DFT}^{-1}(\tilde{S}_{2N}(m,k))$.

This filtering includes a part of the frames before and after the current frame of s, whereas the targeted $|\tilde{S}(m,k)|_0$ was computed only from the current frame of s. As a consequence, the actual values of $|\tilde{S}(m,k)|$ are slightly different from the foreseen values. But taking directly the IFFT of the targeted $|\tilde{S}(m,k)|_0$ would have lead to an undesirable circular convolution.

This synthesis of the sparsified signal $\tilde{s}$ by disjoint blocks may however lead to clicks at the inter-blocks transitions, due to the change of filter coefficients. This phenomenon is particularly noticeable for signals with powerful low-frequency components and a high sampling frequency. For this reason, we smooth the inter-frames transitions by synthesizing $\tilde{s}$ with slightly overlapping blocks.

2.4. Results: sparsity of the sparsified signals

The sparsification described above was tested for a target $\beta' = \beta/2$ on various speech and music signals. We will present here the results for speech, for which the algorithm could be run on a large database. The corpus is composed of 96 source signals from the TIMIT database [11], each consisting in 3 sentences pronounced by the same speaker (96 different speakers), sampled at 16 kHz, truncated to 5, 2 or 1s. The 96×3 sentences are all different and phonetically balanced.

After running the algorithm, we estimated the form factors of $|\tilde{S}(m,k)|_0$, and $|\tilde{S}(m,k)|$, denoted respectively by $\beta^0_{\tilde{s}}$ and $\beta_{\tilde{s}}$, through the moments method. For each source, $\beta^0_{\tilde{s}} \simeq \beta_{\tilde{s}}$, with an error around 0.01. Fig 2 shows the couples $(\beta, \beta_{\tilde{s}})$ for source durations of 1 and 5 s. For each β, $\beta_{\tilde{s}}$ is clearly greater than the target value $\beta/2$. This may be explained by the fact that the traditional histogram equalization is known to fail in reaching exactly a target distribution, especially with few samples, and by a possible inaccuracy of the method of estimation of the form factor. The first hypothesis is reinforced by the fact that the values of $\beta_{\tilde{s}}$ are higher for 1s than for 5s. However, the goal of reducing the form factor was reached in all cases.

2.5. Results: quality of the sparsified signals

The sparsification consists in a filtering varying frame by frame. A distortion measure should encompass the impairments due both to the frequency response at each frame and to the variability of the frequency response between two successive frames. The best way to measure the audibility of this type of complex distortion would be formal subjective tests, resulting in mean opinion scores (MOS) indicating for each source the perceived degradation of the sparsified signal compared to the original one.

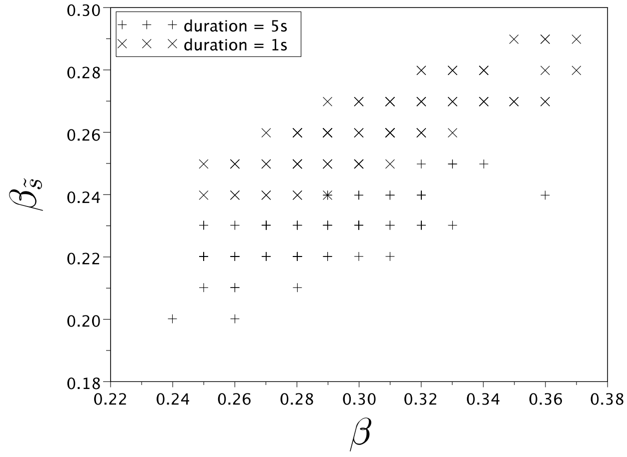


Fig. 2. Estimated form factor $\beta_{\tilde{s}}$ of the sparsified signal *vs* estimated form factor β of the original signal, for 96 sources.

At this stage of the study, we estimated the MOS through PESQ [12], for the 96 previous sources. PESQ provides scores between 1 (very annoying impairment) and 4.5 (no perceptible impairment). The histograms of the estimated MOS, for durations of 1 and 5 s, are represented on Fig. 3. They show that the distortion due to the sparsification is almost inaudible for all the sources for a duration of 5 s. If the duration is shorter, the few number of samples may result in an original distribution of $|S(m,k)|$ far from the Generalized Gaussian hypothesis, which implies more distortion of the signal to reach the target Generalized Gaussian distribution.

3. SEPARATION OF THE SPARSIFIED SIGNALS

In SCA approaches, source separation techniques are usually divided in three steps: (i) identification of the number of sources in the mixtures; (ii) identification of the mixing system; (iii) source separation itself. The quality of the last step is directly related to the accuracy of the former. We will now verify the improvement achieved in the first two steps, when using the doping watermark.

To make the comparison, we will used the ICA+SCA based approach, proposed in [13, 14]. One reason for this choice is that this method showed to be less susceptible to overlapping source signals, so it tends not to favor the proposed approach. Let us summarize the ICA+SCA method for a stereo mixing situation:

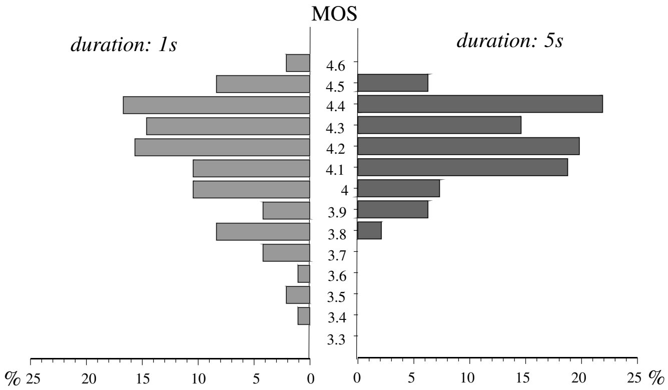


Fig. 3. Histograms of the estimated MOS of the sparsified signals.

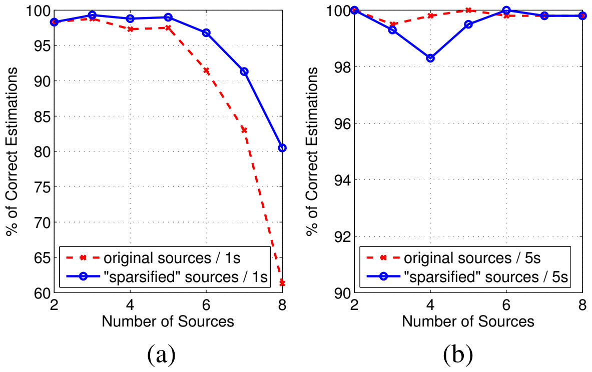
1. Compute FFT of the mixing signals using the same parameters as in the sparsification process.
2. Divide the FFT data in blocks and for each block apply ICA to the mixing signals. The ICA method will provide a ”local separation matrix” $\mathbf{W}_{2\times 2}$.
3. Compute and store all the θ_i obtained by:

$$\theta_i = \tan^{-1}\frac{[\mathbf{W}^{-1}]_{2,i}}{[\mathbf{W}^{-1}]_{1,i}}, \ i = 1, 2 \tag{1}$$

4. Apply K-means [15], or other clustering method in $\boldsymbol{\theta}$, finding the number of clusters that better fits the data. This number will be the amount the sources present in the mixture.
5. The centroid of each cluster will indicate a value of θ that will be related with the direction of one of the columns of the mixing matrix.

3.1. Estimation of the number of sources

The test was run choosing sources randomly among the previous 96 sources. The number of sources varied between 2 and 8, and the mixture was stereo. 400 simulations were done. The results are shown in Fig. 4. For samples of 5 seconds, more than 98% of correct estimations are obtained in both cases. In the case of samples of 1 second, we verified an improvement of the use of the sparsified sources with the grow of the number of sources in the mixture. These results show that is possible to use a smaller amount of samples for the estimation, therefore decreasing the time consumption of this part of the process.



(a)

(b)

Fig. 4. Percentage of correct estimations of the number of sources in the mixture. Average of the results of 400 simulations. **a**: case with 1s sources. **b**: case with 5s sources.

3.2. Estimation of the mixing matrices

The same test was run, now considering that the number of sources is known. In this case, the comparison concerns the quality of the estimation of the mixing matrix. The results are shown using the average values of 200 simulations for each number of sources. The Angular Mean Error (AME) was calculated between the directions of the columns of the estimated and the original mixing matrices. Both the results with 5 and 1 second (see Fig. 5), show better estimations for the sparsified source.

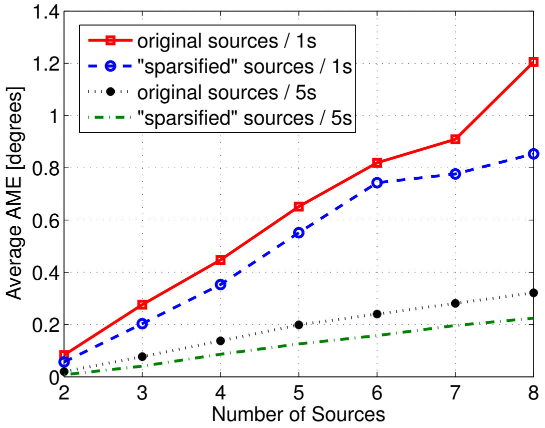


Fig. 5. Average Angular Mean Error (AME) of the directions of the columns of the mixing matrices, given in degrees. Average of the results of 200 simulations.

4. CONCLUSION

We have shown that it is possible to increase the sparsity of the time-frequency representation of an audio signal, through a simple histogram equalization, that preserve the audio quality of the signal. A source separation method based on ICA+SCA have exhibited better performance with "sparsified" signals than with the original ones, in terms of identification of the number of sources and of the mixing system. The next steps of this work will be to explore to which extent the sources can be "sparsified" with respect to the inaudibility constraint and to complete the source separation process until the separation itself. This could lead to applications in "active listening", where a listener can act separately on each source of a record.

At this stage, the proposed method cannot be compared to the ones described in the introduction in terms of separation efficiency, but we can already note two advantages : 1) it is not limited by the watermarking rate constraint; 2) it is not specificaly dedicated to a particular mixing matrix, since the source signals are sparsified for any mix.

Many BSS techniques depend on the distributions of the source signals. Beyond the case presented here, assuming however that the sources are available, the proposed doping watermarking opens many possibilities of controling the distribution of the source signals and, thus, controling the behaviour of various BSS algorithms.

5. ACKNOWLEDGMENTS

We thank Ricardo Suyama, Leonardo T. Duarte and Romis Attux for the fruitful discussions about source separation. Thank you to the Conselho Nacional de Desenvolvimento Científico e Tecnológico, who partly financed this work.

6. REFERENCES

[1] P. Comon and P. Jutten, *Handbook of Blind Source Separation*, Academic Press, 2010.

[2] S. Rickard, "Sparse sources are separated sources," in *Proceedings of the 14th Annual European Signal Processing Conference*, Florence, Italy, September 2006.

[3] M. Parvaix, L. Girin, and JM Brossier, "A Watermarking-Based Method for Informed Source Separation of Audio Signals with a Single Sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1464–1475, Aug 2010.

[4] M. Parvaix, L. Girin, and JM Brossier, "Informed Source Separation of Linear Instantaneous Under-Determined Audio Mixtures by Source Index Embedding ," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721 – 1733, Aug. 2011.

[5] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, September 2010, pp. 498–505.

[6] S. D. Larbi and M. Jaidane, "Audio Watermarking : A Way to Stationnarize Audio Signals," *IEEE Trans. Signal Processing, Supplement on Secure Media*, vol. 53, no. 2, pp. 816–823, February 2005.

[7] I. Marrakchi, G. Mahé, M. Jaidane-Saidane, S. Djaziri-Larbi, and M. Turki-Hadj Alouane, "Gaussianization method for identification of memoryless nonlinear audio systems," in *Proceedings of the European Signal Processing Conference*, 2007, pp. 2316–2320.

[8] H. Halalchi, G. Mahé, and M. Jaidane, "Revisiting quantization theorem through audiowatermarking," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2009, pp. 3361–3364.

[9] J. Pinel and L. Girin, ""'sparsification" of audio signals using the MDCT/IntMDCT and a psychoacoustic model – application to informed audio source separation," in *Proc. of the 42nd Audio Engineering Society Conference: Semantic Audio*, Ilmenau, Germany, 2011.

[10] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1404–1415, October 1989.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993.

[12] "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001, ITU-T Rec. P.862.

[13] E. Z. Nadalin, R. Suyama, and R. Attux, "An ICA-based method for blind source separation in sparse domains," in *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA2009)*, Paraty, Brazil, March 2009, pp. 597–604.

[14] E. Z. Nadalin, R. Suyama, and R. Attux, "Estimating the number of audio sources in a stereophonic instantaneous mixture," in *Proceedings of 7o Congresso de Engenharia de Áudio - AES2009*, May 2009.

[15] C. Chinrungrueng and C. Sequin, "Optimal adaptive k-means algorithm with dynamic adjustment of learning rate," *IEEE Transaction on Neural Networks*, vol. 6, no. 1, pp. 157–169, 1995.