ADAPTATION OF SOURCE-SPECIFIC DICTIONARIES IN NON-NEGATIVE MATRIX FACTORIZATION FOR SOURCE SEPARATION

Xabier Jaureguiberry, Pierre Leveau, Simon Maller, Juan José Burred

Audionamix
114, avenue de Flandre
75019 Paris, France
{firstname.{middlename.}lastname}@audionamix.com

ABSTRACT

This paper concerns the adaptation of spectrum dictionaries in audio source separation with supervised learning. Supposing that samples of the audio sources to separate are available, a filter adaptation in the frequency domain is proposed in the context of Non-Negative Matrix Factorization with the Itakura-Saito divergence. The algorithm is able to retrieve the acoustical filter applied to the sources with a good accuracy, and demonstrates significantly higher performances on separation tasks when compared with the non-adaptive model.

Index Terms— audio source separation, dictionary, non-negative matrix factorization, adaptation

1. INTRODUCTION

Audio source separation is a topic that receives a lot of attention nowadays. Fully automatic source separation is still out of reach, but a number of applications involving a human operator are starting to yield satisfactory results. It is gradually included in software used by sound engineers, and starts to be used for remixing, remastering, upmixing and denoising.

Audio source separation is made possible by exploiting various features that make several sources distinguishable from each other: spatial diversity, spectral shape characteristics, perceptual grouping (Computational Auditory Scene Analysis (CASA)-based methods). When only one channel is available, there is no spatial information to take into account. As a consequence, the two latter features have to be exploited.

In this paper, we address audio source separation based on dictionary learning using spectral shape characteristics. We suppose that there are isolated samples available for some sources involved in the mixture. While most of the existing dictionary-based methods rely on the hypothesis that the training samples have to include samples recorded in similar conditions than in their instantiation on the mixture, our approach tolerates a difference of equalization between the learning source and the source in the mixture. The presented method involves a filter adaptation, which we implemented in a Non-Negative Matrix Factorization (NMF) framework based on the Itakura-Saito (IS) divergence [1].

Section 2 will describe the signal model that was chosen. Then, Section 3 will present the algorithm that performs the source separation. In Section 4, the experimental settings will be detailed and the results will be discussed.

2. SIGNAL MODEL

2.1. Framework

In this study, the problem of source separation is formulated in an NMF framework, with the IS divergence as the metric [1]. More generally, this work can be ascribed to the variance modeling paradigm [2]. We address here the case of mono signals.

The input signal s is transformed into the time-frequency domain by means of a Short Time Fourier Transform (STFT), yielding a matrix $\mathbf{S}$. As in [1], the squared modulus of each element is computed to obtain a matrix of variances $\mathbf{V}$. The problem of NMF is to find the matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{V} \simeq \mathbf{WH}. \tag{1}$$

$\mathbf{W}$ and $\mathbf{H}$ have dimensions $F \times K$ and $K \times N$, and it is desirable that $F \times K + K \times N \ll FN$.

The factorization is formulated as the minimization problem

$$(\mathbf{W}, \mathbf{H}) = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}}\ D_{IS}(\mathbf{V}|\mathbf{WH}), \tag{2}$$

where D_{IS} is a cost function involving the IS divergence d_{IS}:

$$D_{IS}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_{IS}(\mathbf{V}_{(f,n)}|[\mathbf{WH}]_{(f,n)}). \tag{3}$$

The IS divergence is defined as:

$$d_{IS}(x, y) = \frac{x}{y} - \log\frac{x}{y} - 1. \tag{4}$$

This divergence is a good measure for the perceptual difference between two signals, which is explained by its scale invariance: $d_{IS}(\gamma x|\gamma y) = d_{IS}(x|y)$, for a given scalar γ.

The matrix $\mathbf{W}$ obtained after an approximation following this model contains Power Spectral Densities (PSDs) and is commonly called dictionary, whereas $\mathbf{H}$ contains activations of these PSDs across time. If K is carefully chosen, the PSDs constitute a good characterization of the audio sources involved in the mixture.

2.2. Fixed dictionary

Fixed dictionaries for source separation have been used in several previous approaches [3, 4, 5]. In an NMF context, it is possible to fix a number of the columns of $\mathbf{W}$ using prior learning [6]. The learning consists in performing an NMF on an isolated sample of the source that is playing. This process gives a dictionary of PSDs

that can be used as a dictionary for the extraction. However, when the recording conditions of the mixture differ from the ones of the learning sources, the performance of such an approach is limited. This problem is addressed here with a filter adaptation strategy.

2.3. Filter on fixed dictionary

Adaptation of a linear filter applied to a dictionary is a relevant feature for source separation with fixed dictionaries. Indeed, a large part of the differences between various recording conditions and various instruments of a given class can be represented by an acoustical filter. Adapting an acoustical filter for source separation with fixed dictionaries has been proposed in [7] in the context of Gaussian Mixture modeling (GMM) of source PSDs. A similar approach has been developed in [8] where filters are learned to model convolutive mixtures prior to the estimation of activations $\mathbf{H}$. Here, we investigate the use of acoustical filter adaptation in an NMF with the IS divergence (IS-NMF) to take into account the acoustical differences between learning and separation steps for given source signals. In particular, we will discuss two different ways of learning the dictionary.

In the frequency domain and for a given source i, such a filter $\mathbf{g}_i$ multiplies each PSD of a given source, giving the following new source-wise variance approximation:

$$\mathbf{V}_i \simeq \text{diag}(\mathbf{g}_i)\mathbf{W}_i\mathbf{H}_i, \tag{5}$$

where $\text{diag}(\mathbf{g}_i)$ is a diagonal matrix with $\mathbf{g}_i$ as the diagonal vector.

The total variance $\mathbf{V} = |\mathbf{S}|^{\cdot 2}$ (the $^{\cdot x}$ operator denotes element-wise power) of the mixture being:

$$\mathbf{V} = \sum_{i=1}^{I} \mathbf{V}_i \simeq \mathbf{D} = \sum_{i=1}^{I} \text{diag}(\mathbf{g}_i)\mathbf{W}_i\mathbf{H}_i, \tag{6}$$

where $\mathbf{D}$ is the estimated variance of the signal (sum of all the source variance models).

Using this formalism for every source in the signal, two different usage modes can be devised:

- $\mathbf{W}_i$ is fixed, in which case the filter $\mathbf{g}_i$ can be either a free or a fixed vector with unit components (equivalent to a unit gain, $\text{diag}(\mathbf{g}_i) = \mathbf{I}$).
- $\mathbf{W}_i$ is free, in which case integrating the filter $\mathbf{g}_i$ is of no use since it adds a useless frequency indeterminacy.

These models will be illustrated in section 4.

3. ALGORITHM

3.1. Original algorithm for IS-NMF

To solve the IS-NMF factorization, we choose to work with the multiplicative update algorithm [9]. For a model $\mathbf{V} \simeq \mathbf{WH}$, the update rules are:

$$\mathbf{H} \leftarrow \mathbf{H}. * \frac{\mathbf{W}^T((\mathbf{WH})^{\cdot -2}. * \mathbf{V})}{\mathbf{W}^T(\mathbf{WH})^{\cdot -1}} \tag{7}$$

and

$$\mathbf{W} \leftarrow \mathbf{W}. * \frac{((\mathbf{WH})^{\cdot -2}. * \mathbf{V})\mathbf{H}^T}{(\mathbf{WH})^{\cdot -1}\mathbf{H}^T} \tag{8}$$

where $.*$ denotes element-wise multiplication, and $\frac{\mathbf{A}}{\mathbf{B}}$ denotes element-wise division. These update rules are iterated until a maximum number of iterations N_{it} is reached.

3.2. Algorithm for fixed dictionary and free filter

The proposed algorithm is a multiplicative gradient algorithm. The update rule for $\mathbf{H}_i$ is the following:

$$\mathbf{H}_i \leftarrow \mathbf{H}_i. * \frac{(\text{diag}(\mathbf{g}_i)\mathbf{W}_i)^T\mathbf{D}^{\cdot -2}. * \mathbf{V}}{(\text{diag}(\mathbf{g}_i)\mathbf{W}_i)^T\mathbf{D}^{\cdot -1}}. \tag{9}$$

To obtain the update rule for $\mathbf{g}_i$, the cost function has to be derived as a function of this variable. Then, the ratio between the positive part P and the negative part Q of this derivative is computed, then multiplied by the previous value of $\mathbf{g}_i$. This yields:

$$P(f) = \sum_n (\mathbf{V}_{(f,n)}. * [\mathbf{W}_i\mathbf{H}_i]_{(f,n)}./\mathbf{D}^{.2}_{(f,n)}) \tag{10}$$

$$Q(f) = \sum_n ([\mathbf{W}_i\mathbf{H}_i]_{(f,n)}./\mathbf{D}_{(f,n)}) \tag{11}$$

$$\mathbf{g}_i \leftarrow \mathbf{g}_i. * (P./Q) \tag{12}$$

where $./$ denotes the element-wise division. To avoid scale indeterminacies, the filter is normalized to unit energy, the gain information being stored in $\mathbf{H}$.

Considering several sources than can have an adapted filter or not, and a free PSD or a fixed learned one, the problem can be solved with Algorithm 1.

Algorithm 1 IS-NMF with source-specific filter adaptation

Require: $\mathbf{V}$, IsTheFilterToUpdate_i $\forall i$, IsTheDictionaryToUpdate_i $\forall i$

```
% Initialize the matrices for each source
for i = 1 to i = I do
    initialize the filter g_i with ones
    initialize the PSDs W_i with learning or with random values
    initialize the activations H_i with random values
end for
% Perform the factorization
for n_it = 1 to n_it = N_it do
    for i = 1 to i = I do
        if IsTheFilterToUpdate_i then
            g_i ← g_i. * (Σ_t(V.*(W_i H_i)./D^.2)) / (Σ_t(W_i H_i./D))
        end if
        if IsTheDictionaryToUpdate_i then
            W_i ← W_i. * ((D^.-2.*V) H_i^T) / (D^.-1 H_i^T)   % here diag(g_i) = I
        end if
        H_i ← H_i. * ((diag(g_i) W_i)^T D^.-2.*V) / ((diag(g_i) W_i)^T D^.-1)
    end for
end for
return g_i, W_i, H_i
```

3.3. Learning

As stated in the previous section, the learning step can be done with several methods. In this study, we investigate two ways of learning a dictionary: one consists in performing an IS-NMF factorization with a free $\mathbf{W}$ (see Sect. 3.1) while the other is based on a K-means algorithm. The K-means algorithm performs a vector quantization of the squared modulus of the STFT of an input signal. It can be parameterized by K, the number of PSDs that will be kept and stored in the $\mathbf{W}_i$ matrix for each source. For consistency, the LBG K-means method [10] is used, since it relies on the IS divergence.

3.4. Separation

Once each model of the source is computed, Wiener masks $\mathbf{M}_i$ are computed and applied to the mixture spectrogram $\mathbf{S}$ to obtain the source-specific spectrograms $\mathbf{S}_i$:

$$\mathbf{S}_i = \mathbf{M}_i . * \mathbf{S} = (\text{diag}(\mathbf{g}_i)\mathbf{W}_i\mathbf{H}_i)./\mathbf{D} . * \mathbf{S}. \quad (13)$$

The temporal signals s_i of each source are then computed by performing an overlap-add operation from $\mathbf{S}_i$.

4. EXPERIMENTS

This section presents the results obtained with the proposed system. We first validate the signal model with synthetic data, then we show results on real musical instruments. All measurements were obtained with the BSS_EVAL toolbox [11]. Separation sound examples are available online[1].

4.1. Validation

The goal of this step is to validate the filter which is estimated thanks to the update rule (8). Our synthetic database is composed of two pieces of music with original separated tracks. Beforehand, one of the tracks of each piece has been filtered with different types of designed filters (a multi-notch filter and a high-pass one), then mixed with its accompaniment made of all other separated tracks. The test aims at isolating the filtered track from the accompaniment thanks to two source models:

- one for the filtered track, which consists of a fixed dictionary previously learned from the non-filtered corresponding track and a filter to be estimated,
- another for the rest of the mix, which consists of a fixed dictionary learned from the original accompaniment track and without filter. In these experiments, the number of components for the accompaniment model is fixed to $K = 45$.

Once the separation is computed, the estimated filter can be compared to the initially designed filter. For instance, Figure 1 shows the multi-notch filter estimated with a prior dictionary learned with the IS-NMF method. Most of the frequencies are correctly estimated, except in low frequencies where a lack of data causes a mis-estimation.

All experiments have been realized with different numbers of components K for the filtered source. Table 1 gives the best results obtained with the specified value of K. These results highlight that the IS-NMF learning method always reaches better SDR (Signal To Distortion Rate), SIR (Signal to Interference Rate) and SAR (Signal to Artifact Rate) than the K-means one. Furthermore, concerning experiments with IS-NMF learning, the filter estimation always improves the separation quality (up to 6.9 dB) whereas with K-means learning, the improvement is lower and not systematic.

4.2. Real tests

We choose two different classes of instruments to test our approach: two polyphonic instruments, piano and guitar, and one monophonic instrument, bass[2]. The tracks come from real multi-track recordings, so the instruments are expected to play in synchrony and in harmony. Let us note $\mathcal{S} = \{\text{piano, guitar, bass}\}$ the set of sources to be studied. The training signal for each source of $\mathcal{S}$ is built from samples of the RWC database [12], and consists of a concatenation of all the whole range of notes of one single instrument per source. The test data is taken from a recording from which the separated tracks have been made available. For each source of $\mathcal{S}$, we generate a two-source mono mixture which contains the source to separate and another available source (drums). This leads to three different tests :

1. Piano test. Source 1 : piano, source 2 : drums.
2. Guitar test. Source 1 : guitar, source 2 : drums.
3. Bass test. Source 1 : bass, source 2 : drums.

For each test, the learning stage is done on the source-specific training signal for source 1 ($\mathbf{W}_1$ is fixed). Source 2 is modeled with free PSD components ($\mathbf{W}_2$ is updated). Each mixture is 10 s long.

4.2.1. Influence of the learning method

Again, two strategies have been studied for the learning of source-specific dictionaries $\mathbf{W}_1$, the LBG K-means one and the IS-NMF

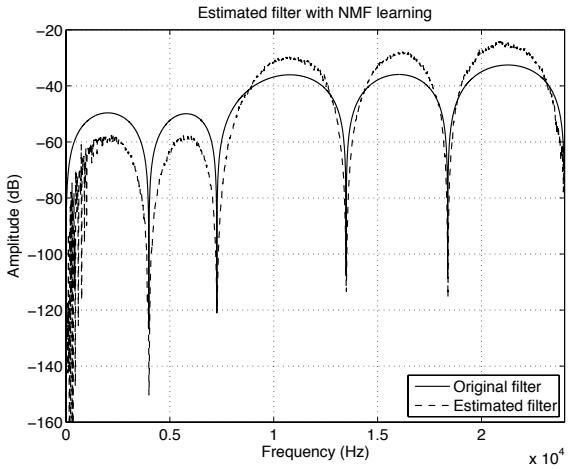

Fig. 1. Example of an estimated filter with IS-NMF-based learning

Filter	Method		SDR	SIR	SAR
SOURCE 1 : Filtered drums / SOURCE 2 : Guitar					
Multi-notch	NMF (K=20)	w/o filter	6.6	16.4	7.2
		w filter	12.9	27.5	13.1
	K-means (K=10)	w/o filter	-4.1	-2.3	5.0
		w filter	7.1	12.9	8.6
High-pass	NMF (K=20)	w/o filter	5.8	12.9	6.9
		w filter	12.7	26.7	12.9
	K-means (K=10)	w/o filter	-6.6	-4.2	2.7
		w filter	7.0	13.4	8.4
SOURCE 1 : Filtered strings / SOURCE 2 : Guitar and drums					
Multi-notch	NMF (K=10)	w/o filter	2.3	9.7	3.6
		w filter	3.3	11.4	4.3
	K-means (K=5)	w/o filter	-3.5	1.7	0.4
		w filter	-8.4	-6.3	3.1
High-pass	NMF (K=10)	w/o filter	0.1	6.6	2.0
		w filter	6.1	16.0	6.6
	K-means (K=5)	w/o filter	-13.0	-10.7	1.9
		w filter	-8.4	-4.5	-0.4

Table 1. Results of separation of a filtered source from its accompaniment

[1] `http://research.audionamix.com/ssd_icassp2011`

[2] Bass can be polyphonic but we only address its monophonic usage here.

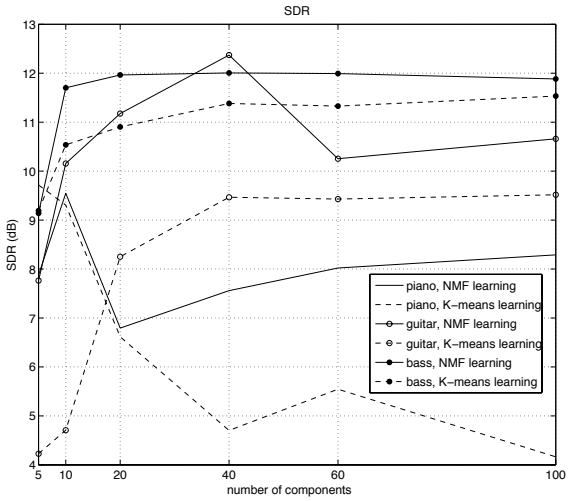

Fig. 2. Evaluation of the two learning methods: IS-NMF and LBG K-means

Experiment		learning	SDR	SIR	SAR
piano	w filter	K-means (K=5)	9.72	11.38	15.12
	w/o filter	NMF (K=100)	-0.08	4.42	3.16
guitar	w filter	NMF (K=40)	12.37	13.74	18.21
	w/o filter	NMF (K=60)	10.75	13.02	14.89
bass	w filter	NMF (K=10)	11.70	12.76	18.59
	w/o filter	NMF (K=60)	9.17	11.62	13.11

Table 2. Best separation results (in dB) for each experiment

one. For both methods, we vary the number of components, K. For each source of $\mathcal{S}$, we have compared the performance of each learning method in the context of IS-NMF with source-specific filter adaptation. The SDR, SIR and SAR measurements as a function of K have been computed for each experiment. For the three criteria, and for each source of $\mathcal{S}$, IS-NMF-based learning seems to better fit our model in most cases. It is illustrated in Figure 2, where the SDR criterion is presented. Note that for small values of K, the two methods give similar results, K-means being more efficient in the case of the piano, for example.

4.2.2. Influence of the filter adaptation

The influence of our approach is studied by comparing the results with and without filter adaptation. A Hamming window of 85 ms with an overlap factor of 75% is used to compute the STFT of the mixtures. The number of points is 4096 with a sampling rate of 44.1 kHz. The secondary source is set at $K = 45$ components. The results are given in Table 2. For each experiment, the learning condition (method and number of components used) which gave the best SDR is indicated and the corresponding SDR, SIR and SAR are shown. Source separation performance is always better when adaptation is performed, with an average gain of 4.9 dB over all experiments and for all criteria. This motivates the use of IS-NMF with source-specific filter adaptation for the three types of sources. An interesting feature of this adaptation is that fewer PSD components are required to decompose a given source adequately. For guitar, only 40 components are necessary to achieve the best separation when an adaptive filter is used, compared to 60 fixed ones, and for bass it is 10 compared to 60. An adapted-filtered dictionary can be both more flexible and more compact than a fixed dictionary.

5. CONCLUSION

In this paper, a filter adaptation of fixed dictionaries in an IS-NMF framework has been proposed. We demonstrated that this adaptation improves the separation quality over the fixed dictionary approach. Future work will consist in deriving smoothing strategies for the filters, in order to prevent them to take inconsistent values for some frequencies. This can occur when there is a lack of data for a given frequency to reliably estimate the filter, and this can lead to local separation artifacts.

6. REFERENCES

[1] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[2] E. Vincent, M.G. Jafari, S.A. Abdallah, M.D. Plumbley, and M.E. Davies, "Probabilistic modeling paradigms for audio source separation," *W. Wang, ed.: Machine Audition: Principles, Algorithms and Systems*, pp. 162–185, 2010.

[3] S.T. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, vol. 13, pp. 793–799, 2001, MIT Press.

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

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and Semi-supervised Separation of Sounds from Single-Channel Mixtures," in *Proc. of Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA)*. 2007, p. 414, Springer.

[6] M.N. Schmidt and R.K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*. Interspeech, 2006.

[7] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[9] I. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," *Advances in neural information processing systems*, vol. 18, pp. 283, 2006.

[10] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.

[11] Vincent E., Gribonval R., and Fvotte C., "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, 2006.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Baltimore, USA, 2003.