# AUDIO SOURCE SEPARATION USING MULTIPLE DEFORMED REFERENCES

*Nathan Souviraà-Labastie*[1]*, *Anaik Olivero*[2], *Emmanuel Vincent*[3], *Frédéric Bimbot*[4]

[1] Université de Rennes 1, IRISA - UMR 6074, Campus de Beaulieu 35042 Rennes cedex, France
[2] Inria, Centre Rennes - Bretagne-Atlantique, 35042 Rennes cedex, France.
[3] Inria, Centre de Nancy - Grand Est, 54600 Villers-lès-Nancy, France
[4] CNRS, IRISA - UMR 6074, Campus de Beaulieu 35042 Rennes cedex, France

## ABSTRACT

This paper deals with audio source separation guided by multiple audio references. We present a general framework where additional audio references for one or more sources of a given mixture are available. Each audio reference is another mixture which is supposed to contain at least one source similar to one of the target sources. Deformations between the sources of interest and their references are modeled in a general manner. A nonnegative matrix co-factorization algorithm is used which allows sharing of information between the considered mixtures. We run our algorithm on music plus voice mixtures with music and/or voice references. Applied on movies and TV series data, our algorithm improves the signal-to-distortion ratio (SDR) of the sources with the lowest intensity by 9 to 12 decibels with respect to original mixture.

***Index Terms*—** Guided audio source separation, nonnegative matrix co-factorization

## 1. INTRODUCTION

Source separation is a cross-cutting field of research dealing with various types of problems and data. This field is rapidly growing taking advantage of its inherent diversity and progress made in each of its components. In the case of audio source separation, achieving the natural human ability of hearing and describing auditory scenes still remains a far end goal. Many approaches have been investigated such as Nonnegative Matrix Factorization (NMF) [1], sparse representations [2] and others.

Blind source separation is an ill-posed problem, and a key point is to embed a maximum amount of a priori information about the sources to guide the separation process [3, 4]. For instance, the general framework presented in [5] proposes to take into account spatial and spectral information about the sources. More recently, a number of approaches have been proposed to exploit information about the recording conditions, the musical score [6], the fundamental frequency $f_0$ [7], the language model [8], the text pronounced by a speaker [9], or a similar audio signal [6, 9–11]. We focus on the latter category of approach where additional information comes from an extra signal called reference.

In [6], the authors propose a model for piano spectrogram restoration, based on generalized coupled tensor factorization where additional information comes from an approximate musical score and spectra of isolated piano sounds. The framework described in [9] proposed a separation model between voice and background guided by another speech sample corresponding to the pronunciation of the same sentence. The speech reference is either recorded by a human speaker or is created with a voice synthesizer based on the available text pronounced by the speaker of the mixture to be separated. A nonnegative matrix co-factorization (NMcF) model is designed so that some of the factorized matrices are shared by the mixture and the speech reference. The authors of [7] incorporate knowledge of the fundamental frequency $f_0$ in a NMF model, by fixing the source part of a source-filter model to be a harmonic spectral comb following the known $f_0$ value of the target source over time. In the context of audio separation of movie soundtracks, the separation can be guided by other available international versions of the same movie [12]. A cover-informed source separation principle is introduced in [10] where the authors assume that cover multitrack signals are available and are used as initialization of the NMF algorithm. The original mixture and the cover signals are time-aligned in a pre-processing step.

In this paper we propose a general framework for *reference-based source separation* which enables joint use of multiple deformed references. For most of previously cited approaches, the framework can either express it as a special case or model the kind of information used in a common formalism : reference signals can be directly available and symbolic additional information can either be synthesized or used to initialize the model. For instance, text-informed separation [9], score informed [6], separation by humming [11], cover guided separation [10] are some of them. In our case, we assume that the musical references have been automatically discovered using an algorithm such as [13], in the context of the separation of voice and music in long audio sequence of movies or TV series.

*Work supported by Maia Studio and Bretagne Region scholarship

This paper is organized as follows. We first describe the NMcF model with audio references in Section 2 and we discuss how it generalizes some of the existing approaches. We also provide an example use case for the separation of a music plus voice mixture guided by music or/and voice references, on single channel audio data coming from TV series and movies. Section 3 provides an algorithmic implementation of the general framework, and Section 4 reports experimental results. We conclude in Section 5.

## 2. GENERAL FRAMEWORK

### 2.1. Input representation

The observations are $M$ single-channel audio mixtures $\mathbf{x}^m(t)$ indexed by $m$. We assume that $\mathbf{x}^1(t)$ is the mixture to be separated, and $\mathbf{x}^m(t)$ for $m > 1$ are the references used to guide the separation process. Each mixture $\mathbf{x}^m(t)$ is assumed to be the sum of sources $\mathbf{y}_j(t)$ indexed by $j \in J_m$ :

$$\mathbf{x}^m(t) = \sum_{j \in J_m} \mathbf{y}_j(t) \text{ with } \mathbf{x}^m(t), \mathbf{y}_j(t) \in \mathbb{R}. \quad (1)$$

In the time-frequency domain, equation (1) can be written as :

$$\mathbf{x}^m_{fn} = \sum_{j \in J_m} \mathbf{y}_{j,fn} \text{ with } \mathbf{x}^m_{fn}, \mathbf{y}_{j,fn} \in \mathbb{C}. \quad (2)$$

The power spectrogram of each source $j$ of the mixture $m$ is denoted as $V_j \in \mathbb{R}_+^{F \times N}$, and the mixture spectrum as $V^m = \sum_{j \in J_m} V_j$. Following the general framework in [5], each $V_j$ is split as the product of an excitation spectral power $V_j^e$ and a filter spectral power $V_j^\phi$. The excitation part (resp. the filter part) is decomposed by an NMF separating the spectral content $W_j^e \in \mathbb{R}_+^{F \times D^e}$ (resp. $W_j^\phi \in \mathbb{R}_+^{F \times D^\phi}$) and the temporal content $H_j^e \in \mathbb{R}_+^{D^e \times N}$ (resp. $H_j^\phi \in \mathbb{R}_+^{D^\phi \times N}$). $D^e$ and $D^\phi$ denote the size of the NMF decomposition of the excitation and the filter. The following decomposition holds :

$$V_j = V_j^e \odot V_j^\phi = W_j^e H_j^e \odot W_j^\phi H_j^\phi \quad (3)$$

where $\odot$ denotes the point wise multiplication, and :

- $W_j^e$ aims to capture the pitch of the source (*e.g.*, frequency range of an instrument or a speaker, and harmonicity)
- $H_j^e$ the corresponding temporal activations (*e.g.*, piano roll or $f_0$ track [7, 11])
- $W_j^\phi$ will capture the spectral envelope (*e.g.*, phoneme dictionary in the case of speech sources [9] or spectral information about an instrument such as isolated notes [6])
- $H_j^\phi$ the corresponding temporal activations (*e.g.*, phoneme alignment for speech or instrument timber changes)

As in [5], the matrices $W$ and $H$ can be either *fixed* (*i.e.*, unchanged during the estimation) or *free* (*i.e.*, adapted to the mixture).

### 2.2. Modeling relationships between the sources of different mixtures

As the different mixtures are composed of similar sources, the matrices $W$ and $H$ can in addition be *shared* (*i.e.*, jointly estimated) between two sources $j$ (from mixture $m = 1$) and $j'$ (from a reference mixture $m'$). We model the deformations between those sources by adding transformation matrices $T$ in the corresponding NMF decomposition (*e.g.*, in the case of one *shared* matrix $W_{j'}^e = T_{jj'}^{fe} W_j^e T_{jj'}^{de}$). Matrices $T$ can be either *fixed* or *free* as $W$ and $H$. When all the matrices $W$ and $H$ are *shared*, the relation becomes :

$$V_{j'} = (T_{jj'}^{fe} W_j^e T_{jj'}^{de} H_j^e T_{jj'}^{te}) \odot (T_{jj'}^{f\phi} W_j^\phi T_{jj'}^{d\phi} H_j^\phi T_{jj'}^{t\phi}) \quad (4)$$

where $V_{j'} \in \mathbb{R}_+^{F' \times N'}$, and :

- $T_{jj'}^{fe}$ (resp. $T_{jj'}^{f\phi}$) $\in \mathbb{R}_+^{F' \times F}$ models the frequency deformations of the excitation (resp. filter) such as equalization or frequency shift (resp. changes in vocal tract length [9]). Note that when $F'$ and $F$ are different, this also enables use of different time-frequency representations.
- $T_{jj'}^{de}$ (resp. $T_{jj'}^{d\phi}$) $\in \mathbb{R}_+^{D^e \times D^e}$ (resp. $\in \mathbb{R}_+^{D^\phi \times D^\phi}$) is a dictionary of deformations of the excitation (resp. filter), and can model pitch shifting, (resp. timber correspondence or different dialects).
- $T_{jj'}^{te}$ (resp. $T_{jj'}^{t\phi}$) $\in \mathbb{R}_+^{F' \times F}$ is the temporal deformation of the excitation (resp. filter), and it is used to time-align the signals. Dynamic time warping can be used to initialize such matrices [9], given that $N'$ and $N$ are usually different. It should also be noticed that using matrices $T^t$ will only align the power spectrum of the mixture. Using phase aligned signals is one of our axis of improvements.

### 2.3. Separation guided by speech and/or music references

In the following, we describe with more details one use case of the previously described general framework, that suits the problem of separating speech and music from old recorded single-channel movies and TV series. To guide and enhance the separation, we consider speech or/and music references. The music references discovered using [13] are intrinsically deformed and contain additional sources such as sound effects. Speech references correspond to the same sentences uttered by different speakers without noise.

In the particular setup reported here, speech sources are numbered 1 and 2, music sources 3 and 4 and noise sources 5, and 6. *Fixed* variables are in black ($W_1^e, W_2^e, W_3^e, W_4^e, T_{34}^{t\phi}$). *Free* variables are in green ( $H_1^e$, $H_2^e$, $T_{12}^{f\phi}$, $T_{12}^{t\phi}$, $T_{34}^{te}$, $W_5$, $H_5$, $W_6$, $H_6$). And variables that are both *free* and *shared* are in red or violet ( $W_1^\phi$, $H_1^\phi$, $H_3^e$, $W_3^\phi$, $H_3^\phi$ ). The *fixed* matrices $T$ set to identity are removed from the notations. The mixture to be separated is modeled as :

$$\begin{aligned} V^1 &= V_1 + V_3 + V_5 \quad (5) \\ &= W_1^e H_1^e \odot W_1^\phi H_1^\phi + W_3^e H_3^e \odot W_3^\phi H_3^\phi + W_5 H_5 \end{aligned}$$

#### 2.3.1. A voice reference mixture

The second mixture is composed of the speech reference alone :

$$V^2 = V_2 = W_2^e H_2^e \odot T_{12}^{f\phi} W_1^\phi H_1^\phi T_{12}^{t\phi} \quad (6)$$

$H_1^e$ and $H_2^e$ are estimated separately to model the different intonations between the speakers, whereas the filter matrices $W_1^\phi$ and $H_1^\phi$ are estimated jointly to model similar phonetic content. $T_{12}^{t\phi}$ models the time realignment between the two pronounced sentences. $T_{12}^{f\phi}$ is constrained to be diagonal and it models both the equalization and the speaker's difference. This model is equivalent to the one used in [9].

#### 2.3.2. A music reference mixture

The third mixture is composed of the music reference $V_4$ supposed to be similar to $V_3$, and some noise $V_6$ :

$$V^3 = V_4 + V_6 = W_4^e H_3^e T_{34}^{te} \odot W_3^\phi H_3^\phi T_{34}^{t\phi} + W_6 H_6 \quad (7)$$

$T_{34}^{te}$ and $T_{34}^{t\phi}$ models the time realignment between the two music examples. Surprisingly it appears that keeping the matrix $T_{34}^{t\phi}$ *fixed* yields better results.

#### 2.3.3. Combining references of different kinds

These two reference models can easily be combined in order to jointly use the three mixtures $\mathbf{x}^1$, $\mathbf{x}^2$ and $\mathbf{x}^3$ during the separation process. The common notations enable us to optimize parameters for each reference.

### 2.4. Extensions of our approach

The proposed framework generalizes the state-of-the-art approaches in [6, 9]. In [6], the source reference is composed of isolated notes and can be modeled with a *shared* $W_j^e$, a *free* $H_j^e$ and by setting the product of $W_j^\phi$ and $H_j^\phi$ to a matrix of ones (no excitation-filter model). The approach described in [9] is exactly expressed by (5) and (6).

Our framework can also model the same kind of information used in [6, 7, 9–11]. The separation by humming approach in [11] can be implemented by sharing the excitation part (*i.e.*, $W_j^e$ and $H_j^e$) between the target source and the reference. Symbolic music or speech information can be used after being synthesized as in [9], or directly in the model as in [6,7] by constraining $H_j^e$. Cover-guided separation as in [10] is also possible by aligning the cover and the mixture to be separated using matrices $T^t$. In [10] the cover is used to initialize the sources but not used during the source estimation. We explain below how our framework can model the deformations between the cover and the source of interest and hence enable the use of the cover during the source estimation.

In addition, this framework can model other kind of information and thus lead to new scenarios of use. We here briefly describe some of them that have not be investigated yet to our knowledge :

- using multiple references for a specific source, *i.e.*, several $j'$ for a single $j$. This will leads to more robust separation, especially in the case of references with additive sources (like when references are automatically obtained with an algorithm such as [13]). For instance in the case of multi-speaker source separation, the speech sources can be guided by several references, ideally containing the same words uttered by the same speaker.
- music source separation for a verse guided by an other verse. In that case the speech sources will have a *shared* excitation ($H_j^e$ and $W_j^e$) but a different filter ($H_j^\phi$) over time. The approach is similar to [14] but we consider the voice as a repeated deformed pattern instead of modeling the background music only.
- cover guided music separation with explicit models for the deformations. The change of an instrument or a singer can be modeled by setting matrix $W_j^\phi$ to *not-shared* or adapting matrix $T^{f\phi}$. Covers played in minor/major or in another tone can also be considered by using $T^{de}$ to model note changes or frequency transposition.

## 3. ALGORITHMIC ASPECTS

In this section, we describe a general algorithm based on multiplicative updates (MU) as well as the initialization used for the matrices from the example of subsection 2.3.

### 3.1. Mutiplicative updates

Following [1], the Itakura-Saito NMF (IS-NMF) model is well-adapted to audio data and provides the following cost function :

$$\mathcal{C}(\Theta) = \sum_{m=1}^{M} \sum_{f,n=1}^{F,N} d_{IS}(X_{fn}^m | V_{fn}^m) \quad (8)$$

where $\Theta$ is the set of parameters to be estimated, *i.e.*, matrices $W$, $H$ and $T$ that are not *fixed*. $X^m = [|\mathbf{x}_{fn}^m|^2]_{fn}$ and $V^m$ are respectively the observation and the estimated spectrum and $d_{IS}(a|b) = a/b - log(a/b) - 1$ is the Itakura-Saito divergence. Following a standard NMF algorithm [1], multiplicative updates (MU) are easily derived from (3), (4) and (8). Due to a lack of space, we will just derive the update for two representative examples : a *non-shared free* variable $W_j^e$ (9) and a *shared free* variable $W_j^e$ (10) between source $j$ of mixture $m = 1$ and $J'$ sources $j'$ of mixtures $m' \neq 1$. We can notice that if for a given $j$ the set of $j'$ is empty (4) becomes (3). The final source estimate are obtained using an adaptive wiener filter.

$$W_j^e \leftarrow W_j^e \odot \frac{[V_j^\phi \odot V^{m\cdot[-2]} \odot X^m][H_j^e]^T}{[V_j^\phi \odot V^{m\cdot[-1]}][H_j^e]^T} \tag{9}$$

$$W_j^e \leftarrow W_j^e \odot \frac{[V_j^\phi \odot V^{m\cdot[-2]} \odot X^m][H_j^e]^T + \sum_{j'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V^{m'\cdot[-2]} \odot X^{m'}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T}{[V_j^\phi \odot V^{m\cdot[-1]}][H_j^e]^T + \sum_{j'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V^{m'\cdot[-1]}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T} \tag{10}$$

### 3.2. Initialization

NMF is known to be sensitive to the initialization of the matrices. We also give here some details on our initialization choices for the use case described in subsection 2.3. Let us also note that, as we work with MU, zeros in the parameters remain unchanged over the iterations.

The *fixed* excitation spectral patterns $W_j^e$ for $j = 1, 2, 3, 4$ are a set of harmonic components computed as in [5]. We initialize the synchronization matrices $T_{12}^{t\phi}$, $T_{34}^{te}$, and $T_{34}^{t\phi}$ with Dynamic Time Warping (DTW) [15] matrices computed on MFCC vectors [16] for speech sources and on chroma vectors for music sources. Following [9], we allow the temporal path to vary within an enlarged region around the estimated DTW path. As long as we work with deformed and noisy data (especially for music), we weight this enlarged path by coefficients of the similarity matrix, in order to avoid obvious initialization errors. We invite the reader to refer to [9] and following works for details on this strategy and a discussion on its influence on the results. The spectral transformation matrix $T_{12}^{f\phi}$ is initialized by the identity matrix. Choosing this matrix to be diagonal leads to time-invariant spectral deformations. The others matrices ( $H_1^e$, $H_2^e$, $W_5$, $H_5$, $W_6$, $H_6$, $W_1^\phi$, $H_1^\phi$, $H_3^e$, $W_3^\phi$, $H_3^\phi$) are initialized with random values.

In addition, we perform 10 iterations of the classical IS-NMF with MU on the reference signals (6) and (7) alone, where the *shared* matrices ($W_1^\phi$, $H_1^\phi$, $H_3^e$, $W_3^\phi$, $H_3^\phi$) and the noise parameters ($W_6$, $H_6$) are updated whereas matrices $T$ are not. For both references, this guided initialization leads to better separation results. After those initializations, $W_6$ and $H_6$ are then set once again to random values, we then perform 10 updates of the main NMcF.

## 4. EXPERIMENTS

### 4.1. Data

As our underlying goal is to separate old audio-visual recordings, we generate the mixture and the references signals to depict such situations. The musical samples and the corresponding references are obtained using the algorithm in [13] that allows the discovery of non-exact repetitions in long audio stream, here movies or TV series. The discovered samples are characterized by distortion of the source of interest (rhythm changes, fade in ) and additional sources (mainly sound effects). Speech examples are taken from the database in [17] in which 16 different speakers uttered the same 238 sentences. We keep 4 musical examples and 4 sentences (two female and two male speakers) to generate the mixtures.

We consider two voice-to-music ratio levels : -6 dB (music as foreground and voice as background, and 12 dB (the inverse case). These levels are close to those effectively observed in movies and TV series. We synthesized such examples in order to obtain objective measures for the evaluation and compare our estimated sources with the original ones. Combining those parameters leads to 32 original mixtures $X$[1]. The original mixtures and the references are about eight seconds long and they are sampled at 16 kHz. Some examples are available online[1].

### 4.2. Results

We here analyze the performance obtained for the use case described in subsection 2.3 with 10 iterations for the NMcF decomposition and initialized according to subsection 3.2. Table 1 shows the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR) and the signal-to-artifact ratio (SAR) [18]. Even when the music ground truth is corrupted, the result is still relevant as it gives a lower bound to the actual non-measurable performance. We consider the three cases described in subsection 2.3 respectively corresponding to the combination of the mixtures (5)-(6), (5)-(7), and (5)-(6)-(7).

Bold values indicate the best SDR. As expected, the best results are most often obtained when all available references are used. The improvements can be deduced from the values in Table 1 after subtraction of the original source-mixture ratio, and a quality improvement is observed in almost all cases. For each voice-to-music ratio levels, the best improvement are achieved for the sources with the lowest intensity, *i.e.*, 9 dB for voice and 12 dB for music.

## 5. CONCLUSION AND FUTURE WORKS

In this paper, we presented a general way to use audio information to separate a given mixture. This model is general enough to take different kinds of audio references into account which are possibly deformed in the frequency domain and in the temporal domain. We described in details a voice and music separation example guided by speech and/or music references using this general framework.

[1] http://maia.gforge.inria.fr/demo/eusipco2014.html

| **Voice-to-music ratio levels** | -6 dB | | 12 dB | |
|---|---|---|---|---|
| | **voice** | **music** | **voice** | **music** |
| **Speech reference** | -3.28\|-2.69\|9.98 | 2.61\|22.03\| 5.04 | 8.57 \|15.15\|14.24 | -2.64\|2.23\|5.04 |
| **Music reference** | 1.99 \| 7.14 \|3.00 | **9.64**\|17.21\|11.00 | 6.54 \|25.54\|10.30 | -0.30\|3.66\|3.00 |
| **Speech and music references** | **3.86** \| 8.84 \|5.38 | 7.76\|17.45\| 8.69 | **11.93**\|26.04\|13.16 | **0.34** \|4.05\|2.93 |

**Table 1**. Comparison of separation guided by speech or/and music references in terms of average SDR|SIR|SAR (dB).

Music separation from international versions [12] take advantage of multichannel that is not handled by this framework yet. In the future, we plan to extend our algorithm to the multichannel case following the multiplicative rules described in [19] or [20]. A first perspective of this work is to use an EM-like algorithm.

A more general perspective will be the design of some automatic processes to choose the initializations of the parameters we have to estimate. Our model can also be improved by adding well-chosen constraints on the parameters. For instance, smoothness constraints on the spectral transformation matrices $T^{f\,\phi}_{j'j}$ can help to derive a more relevant spectral deformation between the target sources and the references.

## 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, 2009.

[2] M. D Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.

[3] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. 14th Int. WIAMIS*, Paris, France, 2013.

[4] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Magazine*, 2014.

[5] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE TASLP*, vol. 20, no. 4, pp. 1118 – 1133, 2012.

[6] U Simsekli, Y. Kenan Yilmaz, and A. Taylan Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *Proc. IEEE ICASSP*, Kyoto, Japan, 2012, pp. 5369–5372.

[7] J.L. Durrieu and J.P. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. LVA/ICA*, Tel-Aviv, Israel, 2012, pp. 438–445.

[8] G.J. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Proc. LVA/ICA*, Tel-Aviv, Israel, 2012, pp. 356–363.

[9] L. Le Magoarou, A. Ozerov, and Q.K.N. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *Proc. IEEE ICASSP*, Vancouver, Canada, 2013, pp. 1–6.

[10] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Professionally-produced music separation guided by covers," in *Proc. ISMIR Conf.*, Porto, Portugal, 2012.

[11] P. Smaragdis and G. Mysore, "Separation by humming : User-guided sound extraction from monophonic mixtures," in *Proc. IEEE WASPAA*, New Paltz, NY, 2009, pp. 69 – 72.

[12] A. Liutkus and P. Leveau, "Separation of music+effects sound track from several international versions of the same movie," in *Proc. 128th AES Convention*, 2010.

[13] L. Catanese, N. Souviraà-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot, "MODIS: an audio motif discovery software," in *Show & Tell - Interspeech 2013*, Lyon, France, 2013.

[14] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *IEEE TASLP*, vol. 21, no. 1, pp. 71–82, 2013.

[15] D.P.W. Ellis, "Dynamic time warping in matlab," 2003.

[16] D.P.W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.

[17] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, and F. Bimbot, "BL-Database: A french audiovisual database for speech driven lip animation systems," Rapport de recherche RR-7711, INRIA, 2011.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] A. Ozerov, C. Févotte, R. Blouet, and J.L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE ICASSP*, Prague, Tchèque, République, 2011, pp. 257–260.

[20] Q.K.N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.