Text-informed audio source separation using nonnegative matrix partial co-factorization

Luc Le Magoarou, Alexey Ozerov, Ngoc Duong

2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAMPTON, UK

TEXT-INFORMED AUDIO SOURCE SEPARATION USING NONNEGATIVE MATRIX PARTIAL CO-FACTORIZATION

Luc Le Magoarou, Alexey Ozerov and Ngoc Q. K. Duong

technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{luc.lemagoarou, alexey.ozerov, quang-khanh-ngoc.duong}@technicolor.com

ABSTRACT

We consider a single-channel source separation problem consisting in separating speech from nonstationary background such as music. We introduce a novel approach called *text-informed separation*, where the source separation process is guided by the corresponding textual information. First, given the text, we propose to produce a speech example via either a speech synthesizer or a human. We then use this example to guide source separation and, for that purpose, we introduce a new variant of the nonnegative matrix partial co-factorization (NMPCF) model based on a so called *excitation-filter-channel* speech model. The proposed NMPCF model allows sharing the linguistic information between the example speech and the speech in the mixture. We then derive the corresponding multiplicative update (MU) rules for the parameter estimation. Experimental results over different types of mixtures and speech examples show the effectiveness of the proposed approach.

Index Terms— Informed audio source separation, text information, nonnegative matrix partial co-factorization, source-filter model

1. INTRODUCTION

Under-determined audio source separation, including the single-channel case, is still very challenging due to its highly ill-posed nature [1]. Thus, the use of any auxiliary information about the sources and/or the mixing process would be helpful to regularize the problem. Many so-called *informed* approaches, which go in this direction, have been proposed recently. As examples, *score-informed* approaches rely on musical score to guide the separation of music recordings [2, 3, 4, 5]. Other algorithms, referred as *user-guided* or *user-assisted*, take into account any kind of input provided by user such as "humming" mimicking the source of interest [6, 7], user-selected F0 track [8], or user-annotated source activity patterns [9, 10, 11]. In line with that, there are speech separation systems informed by speaker gender [12], by a corresponding video [13], or even by natural language structure [14]. However, while spoken text corresponding to the speech in the mixture is often available, e.g., subtitles (approximate transcription) for a movie track or script (exact transcription) in the production phase, to the best of the authors' knowledge, none of the existing approaches exploits this information to guide the separation process.

In this paper we introduce a novel approach that exploits textual information to guide the separation in single channel mixtures. This approach is inspired by the synthesis-based score-informed music separation [2, 5] where a symbolic representation of the underlying acoustic events (the score) is used to synthesise audio examples that are further used to guide the separation. More specifically, the available text is used to generate a speech example, *i.e.*, via a speech synthesizer, which shares the common linguistic information with the speech in the mixture. Note that, in contrast to music, where the temporal mismatch between the sources and the score-synthesized examples is usually linear (the tempo may not be the same, but the rhythm is usually maintained), it is often non-linear for speech. Moreover, while the pitches of the same musical notes are usually on the same frequency locations, there is no reason that the pitches of two different speakers would be the same. In order to handle such kind of variations between the latent source and the speech example, we propose a novel variant of the nonnegative matrix partial co-factorization (NMPCF) model [1][4], which is based on a so-called *excitation-filter-channel (EFC)* speech model that is a new extension of the excitation-filter model [16, 17]. This formulation allows to jointly factorize the spectrogram of the speech example and that of the mixture, while sharing between them the common linguistic information and adapting to each signal the information that differs, such as the temporal dynamics, the recording conditions, the speaker's prosody and timber.

Though text-informed source separation has not been considered yet in the existing works, some related approaches should be mentioned. Pedone *et al.* [18] proposed an algorithm for phoneme-level text to audio synchronisation applied to mixtures of speech and background music. This algorithm relies on the nonnegative matrix factorization (NMF)-like framework where the source-filter models of English phonemes are pre-trained. In this light, given that there is a sort of latent speech modeling guided by text, it could be extended as well for text-informed speech separation. However, while this approach was not evaluated in terms of source separation, and it learns the general phoneme models, our method exploits specific phonemes in a speech example, which is probably pronounced in a closer way to the speech in the mixture. By this difference, we believe that the proposed approach potentially brings better separation performance. Using a sound mimicking the one to be extracted from the mixture to guide the separation, Smaragdis *et al.* introduced a so called *Separation by Humming (SbH)* approach based on the probabilistic latent component analysis (PLCA) [6] while FitzGerald [7] reported a similar method based on the NMF. However, while the performance resulted from PLCA [6] and NMF [7] is limited due to the strong variations between the source and the example, our proposed NMPCF framework models those variations explicitly.

The structure of the rest of the paper is as follows. A general workflow of the proposed framework is presented in Section 2. The

[1] NMPCF model [4] is a particular case of a more general generalized coupled tensor factorization (GCTF) model that was used as well for informed source separation [15].

978-1-4799-1180-6/13/$31.00

NMPCF-based modeling of the speech example and the mixture is then described in Section 3, followed by the model parameter estimation via multiplicative update (MU) rules in Section 4. The proposed approach is evaluated in terms of source separation performance and compared with several baseline approaches, including approaches that do not rely on textual information as well as a text-informed SbH approach [6], over different mixing and example production scenarios in Section 5. Finally we conclude in Section 6.

2. GENERAL WORKFLOW

The general workflow of the proposed approach is depicted in Fig. 1. The source separation algorithm takes as input an audio mixture to be separated and a speech example produced from the given text to guide the separation. The source separation block will be described in details in Section 3. Finally speech and background estimates are reconstructed from the estimated parameters via standard Wiener filtering [19].

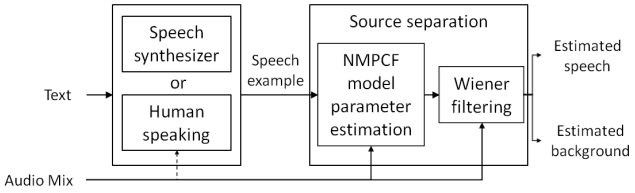

Fig. 1. General workflow of the proposed approach.

One can imagine several ways to generate a speech example that carries the same linguistic information as the speech in the mixture. We identify four main ways to produce such an example. The first way (i) uses the text often provided with TV programs and DVD movies (subtitles or script) to produce the example speech using a speech synthesizer. This scenario is probably among the most attractive ones, since it is totally automatic and does not require any intervention from user. The three other ways we consider are semi-automatic and need the user speaking to produce the example. Depending on the availability of the text and on user's wishes, he/she can either (ii) read the text, (iii) mimic the speech in the mixture after having listened to or (iv) do both.

3. MODELING

Here we describe the proposed NMPCF-based [4] model, as well as the explanation why we chose this specific model. We first formulate the problem, we then describe the mixture and the example models, and finally we introduce the connection between the two models.

3.1. Problem formulation

Let us consider a single-channel mixture:

$$x(t) = s(t) + b(t) \tag{1}$$

consisting of a speech signal $s(t)$ corrupted by a background signal $b(t)$. The goal is to estimate speech, given the mixture $x(t)$ and a speech example $y(t)$.

3.2. Mixture model

Let $\mathbf{X} \in \mathbb{C}^{F\times N}$ be the Short-Time Fourier Transform (STFT) of $x(t)$, F being the number of frequency bins and N the number of time frames. Equation (1) rewrites: $\mathbf{X} = \mathbf{S} + \mathbf{B}$, where $\mathbf{S}$ and $\mathbf{B}$ are the STFTs of the speech and the background, respectively. Defining the power spectrogram $\mathbf{V}_X = |\mathbf{X}|^{\cdot[2]}$ ($\mathbf{A}^{\cdot[b]}$ being the element-wise exponentiation of $\mathbf{A}$ by b), assuming that the speech and background signals are uncorrelated, $\mathbf{V}_X$ can be approximated as:

$$\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \hat{\mathbf{V}}_S + \hat{\mathbf{V}}_B, \tag{2}$$

where $\hat{\mathbf{V}}_X, \hat{\mathbf{V}}_S, \hat{\mathbf{V}}_B \in \mathbb{R}_+^{F\times N}$ are approximations of the power spectrograms of the mixture, the speech and the background, respectively.

We further constrain the speech by imposing a so-called *excitation-filter-channel (EFC)* [2] structure on $\hat{\mathbf{V}}_S$:

$$\hat{\mathbf{V}}_S = \hat{\mathbf{V}}_S^e \odot \hat{\mathbf{V}}_S^\phi \odot \hat{\mathbf{V}}_S^c, \tag{3}$$

with $\odot$ being the Hadamard element-wise product, $\hat{\mathbf{V}}_S^e$ being a time-varying linear combination of comb filters modeling the pitch, $\hat{\mathbf{V}}_S^\phi$ being a time-varying filter modeling the phonemes pronounced, and $\hat{\mathbf{V}}_S^c$ being a time-invariant filter modeling the recording conditions and speaker's vocal tract.

All the matrices in Eq. (3) and matrix $\hat{\mathbf{V}}_B$ are further subject to NMF decompositions as follows:

- $\hat{\mathbf{V}}_S^e = \mathbf{W}^e\mathbf{H}_S^e$, $\mathbf{W}^e \in \mathbb{R}_+^{F\times I}$ being a pre-defined dictionary of combs representing all possible pitches of human voice and $\mathbf{H}_S^e \in \mathbb{R}_+^{I\times N}$ being the corresponding temporal activations.
- $\hat{\mathbf{V}}_S^\phi = \mathbf{W}_S^\phi\mathbf{H}_S^\phi$, $\mathbf{W}_S^\phi \in \mathbb{R}_+^{F\times J}$ being a dictionary of phoneme spectral envelopes and $\mathbf{H}_S^\phi \in \mathbb{R}_+^{J\times N}$ being the corresponding temporal activations.
- $\hat{\mathbf{V}}_S^c = \mathbf{w}_S^c\mathbf{i}_N^T$, $\mathbf{w}_S^c \in \mathbb{R}_+^{F\times 1}$ modeling both the spectral shape of the recording conditions filter and speaker's vocal tract, and $\mathbf{i}_N$ being an N-length column vector of ones.
- $\hat{\mathbf{V}}_B = \mathbf{W}_B\mathbf{H}_B$, $\mathbf{W}_B \in \mathbb{R}_+^{F\times K}$ being a dictionary of background spectral shapes and $\mathbf{H}_B \in \mathbb{R}_+^{K\times N}$ being the corresponding temporal activations.

Another assumption is made so as to constrain spectral shapes of matrices $\mathbf{W}_S^\phi$ and $\mathbf{w}_S^c$ to be smooth [20]. Following [20], these matrices are constrained as follows : $\mathbf{W}_S^\phi = \mathbf{P}\mathbf{E}_S^\phi$ and $\mathbf{w}_S^c = \mathbf{P}\mathbf{e}_S^c$, where $\mathbf{P} \in \mathbb{R}_+^{F\times L}$ is a pre-defined matrix of L so-called *spectral blobs*, that are used to construct $\mathbf{W}_S^\phi$ and $\mathbf{w}_S^c$ with weights $\mathbf{E}_S^\phi \in \mathbb{R}_+^{L\times J}$ and $\mathbf{e}_S^c \in \mathbb{R}_+^{L\times 1}$, respectively.

Finally, the mixture model can be summarized as:

$$\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \overbrace{\underbrace{\left(\mathbf{W}^e\mathbf{H}_S^e\right)}_{\hat{\mathbf{V}}_S^e} \odot \underbrace{\left(\mathbf{W}_S^\phi\mathbf{H}_S^\phi\right)}_{\hat{\mathbf{V}}_S^\phi} \odot \underbrace{\left(\mathbf{w}_S^c\mathbf{i}_N^T\right)}_{\hat{\mathbf{V}}_S^c}}^{\hat{\mathbf{V}}_S} + \overbrace{\mathbf{W}_B\mathbf{H}_B}^{\hat{\mathbf{V}}_B} \tag{4}$$

3.3. Speech example model

Let $\mathbf{Y} \in \mathbb{C}^{F\times N'}$ be the STFT of $y(t)$ and $\mathbf{V}_Y = |\mathbf{Y}|^{\cdot[2]} \in \mathbb{R}_+^{F\times N'}$ its power spectrogram. The example consists of only one clean speech source, whose power spectrogram is approximated as:

$$\mathbf{V}_Y \approx \hat{\mathbf{V}}_Y = \hat{\mathbf{V}}_Y^e \odot \hat{\mathbf{V}}_Y^\phi \odot \hat{\mathbf{V}}_Y^c, \tag{5}$$

[2]The proposed EFC model is an extension of the excitation-filter model [17].

where $\hat{\mathbf{V}}_Y^e$, $\hat{\mathbf{V}}_Y^\phi$ and $\hat{\mathbf{V}}_Y^c$ are decomposed the same way as in Section 3.2, i.e., $\hat{\mathbf{V}}_Y^e = \mathbf{W}^e\mathbf{H}_Y^e$, $\hat{\mathbf{V}}_Y^\phi = \mathbf{W}_Y^\phi\mathbf{H}_Y^\phi$ and $\hat{\mathbf{V}}_Y^c = \mathbf{w}_Y^c\mathbf{i}_{N'}^T$. The smoothness constraints are applied as well: $\mathbf{W}_Y^\phi = \mathbf{P}\mathbf{E}_Y^\phi$ and $\mathbf{w}_Y^c = \mathbf{P}\mathbf{e}_Y^c$.

Finally, the example model can be summarized as:

$$\mathbf{V}_Y \approx \hat{\mathbf{V}}_Y = \underbrace{\left(\mathbf{W}^e\mathbf{H}_Y^e\right)}_{\hat{\mathbf{V}}_Y^e} \odot \underbrace{\left(\mathbf{W}_Y^\phi\mathbf{H}_Y^\phi\right)}_{\hat{\mathbf{V}}_Y^\phi} \odot \underbrace{\left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right)}_{\hat{\mathbf{V}}_Y^c}. \quad (6)$$

3.4. Couplings between the mixture and example models

The role of the example is to guide source separation, thanks to the fact that it shares common linguistic information with the speech in the mixture. We model this sharing as follows:

- The phonemes pronounced in the mixture and those pronounced in the example are the same, thus we assume: $\mathbf{W}_S^\phi = \mathbf{W}_Y^\phi = \mathbf{W}^\phi$, and $\mathbf{W}^\phi$ is to be estimated. This assumption implies $\mathbf{E}_S^\phi = \mathbf{E}_Y^\phi = \mathbf{E}^\phi$.
- The phonemes are pronounced in the same order in the mix and in the example, but not exactly temporally synchronized. Thus we represent $\mathbf{H}_S^\phi$ as $\mathbf{H}_S^\phi = \mathbf{H}_Y^\phi\mathbf{D}$ where $\mathbf{D} \in \mathbb{R}_+^{N'\times N}$ is a so-called *synchronization matrix* [18]. $\mathbf{H}_Y^\phi$ and $\mathbf{D}$ are to be estimated.

Note that these assumptions are reasonable since the mixture and the example contain utterances of the same sentences. The final NMPCF model is as follows:

$$\begin{aligned}\mathbf{V}_Y &\approx \hat{\mathbf{V}}_Y = \left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{W}^\phi\mathbf{H}_Y^\phi\right) \odot \left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right),\\ \mathbf{V}_X &\approx \hat{\mathbf{V}}_X = \left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{W}^\phi\mathbf{H}_Y^\phi\mathbf{D}\right) \odot \left(\mathbf{w}_S^c\mathbf{i}_N^T\right) + \mathbf{W}_B\mathbf{H}_B,\end{aligned} \quad (7)$$

where pre-defined and fixed parameters are $\mathbf{W}^e$, $\mathbf{i}_{N'}^T$ and $\mathbf{i}_N^T$ (in green), shared and estimated parameters are $\mathbf{W}^\phi$ and $\mathbf{H}_Y^\phi$ (in red), and non-shared and estimated parameters are the others (in black). This modeling is visualized on Fig. 2.

To summarize, the parameters to be estimated are [3]:

$$\boldsymbol{\theta} = \left\{\mathbf{H}_Y^e, \mathbf{H}_S^e, \mathbf{E}^\phi, \mathbf{H}_Y^\phi, \mathbf{D}, \mathbf{e}_Y^c, \mathbf{e}_S^c, \mathbf{H}_B, \mathbf{W}_B\right\}, \quad (8)$$

while $\mathbf{W}^e$, $\mathbf{P}$, $\mathbf{i}_N^T$ and $\mathbf{i}_{N'}^T$ are pre-defined and fixed.

4. PARAMETER ESTIMATION

4.1. Cost function

The general principle of NMF-like parameter estimation is to minimize certain cost function measuring a divergence between the data matrix and its structural approximation. We consider here the Itakura-Saito (IS) divergence [4] and specify the cost function as follows:

$$C(\boldsymbol{\theta}) = \lambda \sum_{f,n=1}^{F,N'} d_{IS}\left(v_{Y,fn}|\hat{v}_{Y,fn}\right) + \sum_{f,n=1}^{F,N} d_{IS}\left(v_{X,fn}|\hat{v}_{X,fn}\right), \quad (9)$$

where $\lambda \in \mathbb{R}_+$ is a penalty parameter that determines the example's influence on the estimation, $d_{IS}(a|b) = a/b - \log(a/b) - 1$ is the IS divergence, $v_{Y,fn}$, $v_{X,fn}$, $\hat{v}_{Y,fn}$ and $\hat{v}_{X,fn}$ are, respectively, entries of data matrices $\mathbf{V}_Y$, $\mathbf{V}_X$ and their structural approximations $\hat{\mathbf{V}}_Y$, $\hat{\mathbf{V}}_X$ from (7).

[3] Keep in mind that $\mathbf{W}^\phi = \mathbf{P}\mathbf{E}^\phi$, $\mathbf{w}_S^c = \mathbf{P}\mathbf{e}_S^c$ and $\mathbf{w}_Y^c = \mathbf{P}\mathbf{e}_Y^c$.

[4] When applied to power spectrograms of audio signals, IS divergence was shown as one of the most suitable choices for NMF-like decompositions [19], in particular thanks to its scale invariance property.

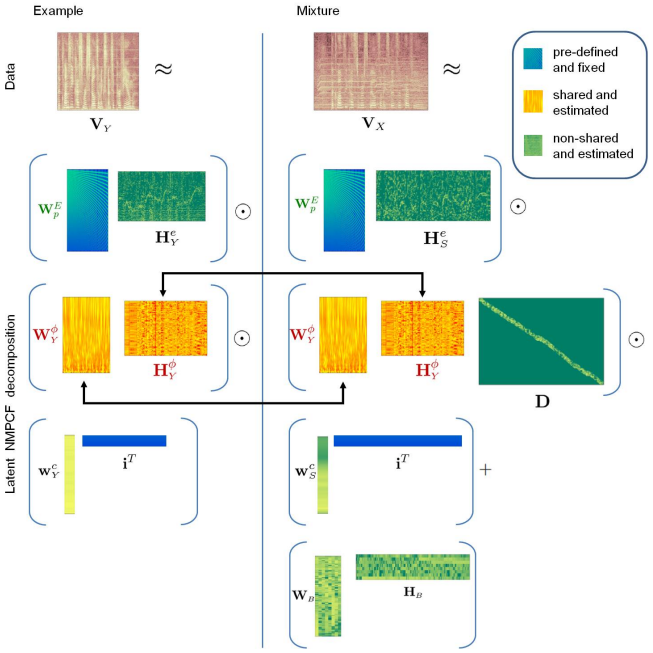

Fig. 2. NMPCF model for the speech example and the mixture.

4.2. Parameter estimation via MU rules

To optimize cost (9) we used standard MU rules which can be derived following a recipe described in [19]. The idea is to derive multiplicative updates based on the cost function's gradient with respect to each parameter. Most of the resulting updates are very similar to those described, e.g., in [20], thus, due to lack of space, we here give only the updates for shared parameters, i.e., $\mathbf{E}^\phi$ and $\mathbf{H}_Y^\phi$ (a complete list of update rules will be given in a longer paper):

$$\mathbf{E}^\phi \leftarrow \mathbf{E}^\phi \odot \frac{\mathbf{P}^T\left[\lambda\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right)\odot\left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right)\odot\hat{\mathbf{V}}_Y^{\cdot[-2]}\odot\mathbf{V}_Y\right)\mathbf{H}_Y^{\phi\,T} + \left(\left(\mathbf{W}^e\mathbf{H}_S^e\right)\odot\left(\mathbf{w}_S^c\mathbf{i}_N^T\right)\odot\hat{\mathbf{V}}_X^{\cdot[-2]}\odot\mathbf{V}_X\right)\mathbf{H}_S^{\phi\,T}\right]}{\mathbf{P}^T\left[\lambda\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right)\odot\left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right)\odot\hat{\mathbf{V}}_Y^{\cdot[-1]}\right)\mathbf{H}_Y^{\phi\,T} + \left(\left(\mathbf{W}^e\mathbf{H}_S^e\right)\odot\left(\mathbf{w}_S^c\mathbf{i}_N^T\right)\odot\hat{\mathbf{V}}_X^{\cdot[-1]}\right)\mathbf{H}_S^{\phi\,T}\right]}, \quad (10)$$

$$\mathbf{H}_Y^\phi \leftarrow \mathbf{H}_Y^\phi \odot \frac{\lambda\mathbf{W}_Y^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right)\odot\left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right)\odot\hat{\mathbf{V}}_Y^{\cdot[-2]}\odot\mathbf{V}_Y\right) + \mathbf{W}_S^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_S^e\right)\odot\left(\mathbf{w}_S^c\mathbf{i}_N^T\right)\odot\hat{\mathbf{V}}_X^{\cdot[-2]}\odot\mathbf{V}_X\right)\mathbf{D}^T}{\lambda\mathbf{W}_Y^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right)\odot\left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right)\odot\hat{\mathbf{V}}_Y^{\cdot[-1]}\right) + \mathbf{W}_S^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_S^e\right)\odot\left(\mathbf{w}_S^c\mathbf{i}_N^T\right)\odot\hat{\mathbf{V}}_X^{\cdot[-1]}\right)\mathbf{D}^T}. \quad (11)$$

Let us just note that these shared parameters updates depend on both data matrices ($\mathbf{V}_Y$ and $\mathbf{V}_X$), on all NMPCF model parameters, as well as on the penalty parameter λ.

5. EXPERIMENTS

In this section we first describe the data, the parameter settings and initialization. We then summarize the baseline methods. Finally, we present and discuss the simulation results.

5.1. Data

We evaluate our approach on the synthetic data which consists of three sets: the mixtures, the speech examples, and the training set needed for some baseline approaches. All audio signals are mono and sampled at 16000 Hz.

The *mixture set* consists of eighty different mixtures created as follows. Ten speech signal (five for male voice and five for female voice) in English corresponding to ten different sentences were randomly selected from the test subset of the TIMIT database [21]. Each chosen speech signal was used to produce eight mixtures by adding to it either music or effect background. These background sounds were extracted from real movie audio tracks and the Signal (*speech*) to Noise (*background*) Ratios (SNRs) were set to four different values: -5dB, 0dB, 5dB and 10dB.

The *example set* is built in accordance to the mixture set. For each of ten TIMIT sentences, twelve corresponding speech examples were created. Two of them (referred below as SYNTH) were produced via speech synthesizers (one with male voice and one with female voice). [5] Other eight examples were produced by human speakers: two by a female native English speaker, two by a male native English speaker, two by a female non-native speaker (Spanish), and two by a male non-native speaker (French). Each of these speakers produced two examples: the first example by just reading the sentence, and another one by reading the sentence, listening to the mixture, and trying to mimic it. These produced examples are referred below as NN-READ for non-native reading, NN-MIM for non-native mimicking, NT-READ for native reading, and NT-MIM for native mimicking. The last two examples (referred below as TIMIT) were taken from the TIMIT test database, but by different speakers: one male and one female. Note that this example set covers three of the four generating scenarios mentioned in Section 2.

The *training set*, which is used only in several baselines, consists of one hundred spoken sentences from different speakers: fifty males and fifty females. These speech signals were randomly selected from the TIMIT train database.

5.2. Parameter setting and initialization

The STFT is computed using a half-overlapping sine window of length 32 ms (*i.e,* 512 samples). Each column of the $F \times I$ excitation dictionary matrix $\mathbf{W}^e$ is set to a harmonic comb with a given fundamental frequency (pitch). The pitch is varied from 100 Hz to 400 Hz covering mostly frequency range of human speech, with an increment of 1/8 of a tone. The entries in the last column of $\mathbf{W}^e$ are set to the same constant value for representing unvoiced sounds. These settings lead to $I = 186$ columns in $\mathbf{W}^e$. The $F \times L$ matrix $\mathbf{P}$ of spectral blobs, which is used to constrain the dictionary of phonemes and the time-invariant channel filters, is built following the auditory-motivated Equivalent Rectangular Bandwidth (ERB) scale [20]. In our experiment L is set to 55. The two matrices $\mathbf{W}^e$ and $\mathbf{P}$ are computed using the Flexible Audio Source Separation Toolbox (FASST) [20] routines. Finally, the penalty parameter λ in (9) is set to $\lambda = N/N'$ so as to compensate for a possible difference between the duration of the example and the mixture.

All parameters in (8) are randomly initialized by positive values, except the synchronization matrix $\mathbf{D}$, which is initialized as follows. Assuming that there is a binary matrix $\mathbf{D}_0 \in \{0,1\}_+^{N' \times N}$ representing an initial synchronization between the example and the speech in the mixture. This matrix is supposed to contain all zero entries except the ones lying in a continuous path connecting the upper left corner and the lower right corner of the matrix as shown in Fig. 3. Note that this path can not go up nor to the left of the current non-zero entry point. We consider two following ways to compute $\mathbf{D}_0$:

- LIN: non-zero path in $\mathbf{D}_0$ follows the straight line connecting upper left and lower right corners (see Fig. 3, left).
- DTW: $\mathbf{D}_0$ is a Dynamic Time Warping (DTW) based alignment matrix between the example and the mixture (see Fig. 3, right). We used the DTW implementation from [22] computed on Mel-Frequency Cepstral Coefficients (MFCCs) with mean and variance normalization, which is known making MFCCs more robust with respect to convolutive and additive perturbations [23].

Since none of the two above mentioned initialization strategies guarantees the perfect synchronization, in particular due to either the speed mismatch between the mixture and the example (LIN case) or the corrupting noise (DTW case), we introduce the initialization strategy for the matrix $\mathbf{D}$ inspired by [18], where we allow the synchronization path to vary within an enlarged region of width B around $\mathbf{D}_0$ as:

$$\mathbf{D} = \mathbf{D}_0 + \mathrm{sign}\left(\sum\nolimits_{b=0}^{(B-1)/2} (\mathbf{U}^b + \mathbf{L}^b)\mathbf{D}_0\right), \qquad (12)$$

where $\mathrm{sign}(\cdot)$ is applied element-wise, and $\mathbf{U}$ and $\mathbf{L} \in \{0,1\}_+^{N' \times N'}$ are, respectively, the upper and the lower shift matrices. Four different initializations of $\mathbf{D}$, which correspond to the LIN case and DTW case with two different values of B ($B = 5$ and $B = 19$), are shown on Fig. 3.

5.3. Baseline approaches

We present below several baseline methods we judge relevant to be compared to.

5.3.1. Baselines non-informed by example

The following approaches use neither speech example nor text information[6]:

- NMF: A standard NMF-based method with a general voice spectral dictionary $\mathbf{W}_S \in \mathbb{R}_+^{F \times J}$ ($J = 128$) which is first learned on the training set described in Section 5.1, and then fixed during the parameter estimation.
- EFC-N: A method using the same EFC mixture model (4) in a non-supervised manner, as in [17], *i.e.*, filter matrices $\mathbf{W}_S^\phi$ and $\mathbf{H}_S^\phi$ are left free and not coupled with example. In other words, this method corresponds to the proposed approach with $\lambda = 0$ in (9).
- EFC-S: A method using the same EFC mixture model (4), which however is not supervised by example any more, but by our training data. In this approach filter dictionary $\mathbf{W}_S^\phi$ is pre-learned on the training set and then fixed during parameter estimation, while $\mathbf{H}_S^\phi$ is updated.

[5] We used "ivona" synthesizers `www.ivona.com/en/` to create speech examples.

[6] We implemented these approaches with help of the FASST [20].

5.3.2. *Baseline informed by example*

We also consider as a baseline the SbH PLCA-based method [6], within the proposed general workflow, as shown on Fig. 1. Since the mixture $\mathbf{V}_X$ and the example $\mathbf{V}_Y$ are not aligned in general, we used $\mathbf{V}'_Y = \mathbf{V}_Y\mathbf{D}_0$ as example for SbH, where $\mathbf{D}_0$ is the initial synchronization computed with DTW as described in Section 5.2. The SbH itself was implemented following [6]. This baseline is referred hereafter as SbH-DTW.

5.4. Simulation results

In this paragraph we first compare the performance of the proposed method with that of the baselines. We then analyze the proposed method's performance variation depending on speech example types.

5.4.1. *Comparison with different baselines*

We consider four variants of the proposed approach referred as LIN19, DTW19, LIN5 and DTW5, which correspond to different synchronization matrix initializations (with LIN or DTW, with $B = 19$ or $B = 5$, see also Fig. 3). We compare the performance achieved by these variants with four baselines described in Section 5.3 on our database. Table 1 shows average results for different mixture types in terms of both the Signal to Distortion Ratio (SDR) criterion [24] and the Overall Perceptual Score (OPS) criterion [25] computed for the speech source only.

We can see in Table 1 that the proposed method gives better average results than all the baselines, and especially on the mixtures with low SNRs (difficult cases). Moreover, in term of the perceptual measure (OPS), the proposed method obtains better results when the synchronization matrix $\mathbf{D}$ is initialized with a wider band ($B = 19$). Finally, DTW initialization leads to slightly better results, as compared to LIN initialization. The best average performance in term of SDR, *i.e.*, $SDR = 6.8$ dB, is obtained by DTW5 initialization. Thus for the experiments in Section 5.4.2, DTW5 initialization strategy is always used to initialize the synchronization matrix.

In order to explicitly assess the importance of textual information to the proposed method, we performed similar experiments with $\mathbf{D}$ initialized by DTW5, where for each mixture the speech example used to guide the separation was randomly chosen to contain a different spoken sentence from the speech in the mixture. The average SDR and OPS achieved in this case are 5.91 dB and 25.9, respectively, which is clearly worse than the average results for DTW5 in Table 1 (average SDR $= 6.80$ dB and average OPS $= 29.4$). This implies that the correctness of the textual information is very important within the considered framework.

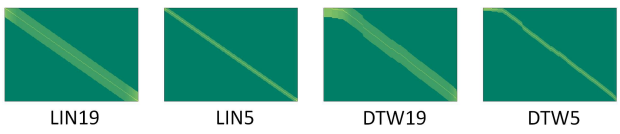

Fig. 3. Different initializations of the synchronization matrix $\mathbf{D}$.

5.4.2. *Performance depending on example types*

In this experiment, we first study the influence of gender disparity between the speech in the mixture and that in the example. For this purpose, we rearrange the results displayed in Table 1 to exhibit the average results depending on the gender disparity between the example's speaker and the mixture's speaker. The results are summarized in Table 2 where the average performances are better, for both two different types of background, in case the speaker of the example and the one of the mixture are of the same gender.

	Music	Effects	Avg
Different gender	6.17 | 28.1	6.92 | 29.5	6.55 | 28.8
Same gender	**6.23** | **29.0**	**7.88** | **31.0**	**7.06** | **30.0**

Table 2. Influence of gender disparity on source separation performance. Results are shown in the form of SDR | OPS measures and the highest value of each column is in bold.

Finally, we study the influence of the example type on the source separation performance. For this purpose, we display the results of Table 1 depending on the example's type. The average results are given in Table 3. One can notice that the performances obtained with the synthesized speech examples and with the real-world recorded ones from human speaking are not very different. Nevertheless, we can see that the examples from TIMIT database provide the best results in all cases. This may be because they have been recorded in better conditions with less noise and reverberation, or because they have been recorded in the same conditions with the speech in the mixtures. It is also observed that, as expected, the mimicked examples give better results than the read ones (at least in terms of the OPS measure) since the mimicked sound matches better the one in the mixture in terms of both temporal variation and spectral patterns. Finally, one can notice a slight improvement in term of the OPS for the examples from native speakers over those from non-native ones.

Example's type	Music	Effects	Avg
SYNTH	5.81 | 28.71	7.07 | 30.16	6.44 | 29.44
NN-READ	5.98 | 26.56	7.32 | 28.14	6.65 | 27.35
NT-READ	6.28 | 28.46	6.85 | 30.18	6.57 | 29.32
NN-MIM	6.00 | 27.41	7.92 | 29.84	6.96 | 28.62
NT-MIM	5.79 | 29.02	6.59 | 31.13	6.19 | 30.07
TIMIT	**7.33** | **30.87**	**8.65** | **32.07**	**7.99** | **31.47**

Table 3. Influence of the example's type on source separation performance. Results are shown in the form of SDR | OPS measure and the highest value of each column is in bold.

The method has been extended to multichannel mixtures and entered into the Signal Separation Evaluation Campaign (SiSEC 2013).

6. CONCLUSION

In this paper, we presented an informed source separation approach that takes into account the available textual information to guide the separation in single channel audio mixtures. We proposed a novel NMPCF modeling framework that allows to efficiently handle the variations between the speech example and the speech in the mixture. The experimental results over different settings confirm the benefit of the proposed approach over both the non-informed NMF-based baseline method, informed NMF-based baseline approaches, and a SbH state-of-the-art algorithm [6]. The proposed approach is experimentally shown not to be very sensitive to the way the speech example is produced. Future work will consist in exploiting some parameters of speech example NMF decomposition as hyper parameters of a prior distribution to guide the separation process.

	Method	Speech + Music				Speech + Effects				Avg
		-5dB	0dB	5dB	10dB	-5dB	0dB	5dB	10dB	
Baselines	NMF	-2.60|20.1	2.22|19.8	7.05|18.7	11.58|20.0	-3.39|26.7	0.66|27.6	4.96|30.4	11.61|33.2	4.01|24.5
	EFC-N	-1.82|24.3	3.07|23.4	8.38|22.9	11.74|22.1	-1.38|24.9	5.15|23.0	**10.74**|23.7	**13.70**|24.4	6.20|23.6
	EFC-S	-2.83|23.1	2.34|22.8	7.19|23.0	11.73|20.8	-3.45|25.1	1.40|26.1	5.95|26.9	10.48|27.5	4.10|24.4
	SbH-DTW	-2.21|10.5	3.40|14.6	6.80|17.4	7.97|21.3	**1.05**|20.6	5.63|27.5	8.21|31.2	9.56|32.0	5.05|21.9
Proposed	LIN19	-1.10|**24.4**	4.21|28.3	8.54|**32.0**	**11.89**|34.1	0.28|26.8	5.27|**30.4**	10.24|33.2	13.24|34.4	6.57|30.4
	DTW19	-1.05|24.3	4.15|**28.6**	8.28|31.7	11.74|**34.7**	-0.02|**27.4**	5.60|30.3	10.43|**33.3**	13.24|**35.0**	6.55|**30.7**
	LIN5	-0.48|22.3	4.41|25.1	8.55|26.8	11.45|28.8	0.18|24.4	5.42|26.9	9.47|28.1	12.30|29.4	6.41|26.5
	DTW5	**-0.38**|24.1	**4.73**|27.1	**8.65**|29.9	11.80|33.0	0.38|26.3	**6.09**|28.9	10.36|32.1	12.77|33.7	**6.80**|29.4

Table 1. Comparison of different configurations of the proposed method with different baselines. Results are shown in the form of SDR|OPS measures and the highest value of each column is in bold.

7. ACKNOWLEDGEMENT

The authors would like to thank S. Ayalde, F. Lefebvre, A. Newson and N. Sabater for their help in producing the speech examples.

8. REFERENCES

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[2] J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders, "Source separation by score synthesis," in *Proc. Int. Computer Music Conference (ICMC)*, New York, NY, June 2010, pp. 462–465.

[3] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 45–48.

[4] U. Simsekli and A. T. Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2639 – 2643.

[5] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing (ICASSP)*, 2013, pp. 888–891.

[6] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proceedings IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[7] D. FitzGerald, "User assisted source separation using non-negative matrix factorisation," in *22nd IET Irish Signals and Systems Conference*, Dublin, 2011.

[8] J.L. Durrieu and J.P. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, March 2012, pp. 438–445.

[9] 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 Int. Conf. on Acoustics, speech, and signal processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 257 – 260.

[10] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, Porto, Portugal, Oct. 2012, pp. 115–120.

[11] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2654–2658.

[12] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13*. 2000, pp. 793–799, MIT Press.

[13] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," in *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing*, Philadelphia, USA, 2005, pp. 425–428.

[14] G. J. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA / ICA)*, Tel-Aviv, Israel, March 2012, pp. 356–363.

[15] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[16] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, 2006.

[17] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[18] A. Pedone, J. J. Burred, S. Maller, and P. Leveau, "Phoneme-level text to audio synchronization on speech signals with background music," in *Proc. INTERSPEECH*, 2011, pp. 433–436.

[19] 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.

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

[21] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT: Acoustic-phonetic continuous speech corpus," Tech. Rep., NIST, 1993, distributed with the TIMIT CD-ROM.

[22] D. Ellis, "Dynamic time warp (DTW) in Matlab," Web resource, 2003, available: `http://www.ee.columbia.edu/ln/labrosa/matlab/dtw/`.

[23] C. P. Chen, J. Bilmes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on Aurora 2.0," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2002, pp. 2445–2448.

[24] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[25] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057.