# SPEECH-GUIDED SOURCE SEPARATION USING A PITCH-ADAPTIVE GUIDE SIGNAL MODEL

*Romain Hennequin, Juan José Burred, Simon Maller and Pierre Leveau*

Audionamix
171 quai de Valmy
75010 Paris, France
< *firstname* >.< *lastname* >@audionamix.com

## ABSTRACT

In this paper, we present a new method to perform underdetermined audio source separation using a spoken or sung reference signal to inform the separation process. This method explicitly models possible differences between the spoken reference and the target signal, such as pitch differences and time lag. We show that the proposed algorithm outperforms state-of-the art methods.

***Index Terms—*** Audio source separation, non-negative matrix factorization, informed source separation.

## 1. INTRODUCTION

Underdetermined source separation has been a key topic in audio signal processing for the last decade. It consists in isolating different meaningful parts in the sound, such as isolating the lead vocal from the accompaniment in a piece of music or the dialog from the background music and effects in a movie soundtrack.

The problem has often been addressed within a blind source separation context using Non-negative Matrix Factorization (NMF) [1]. One of the main drawbacks of this technique is the difficulty to cluster the factorized elements and associate them with a source.

Recently, numerous works proposed to add extra information in order to improve separation results.

Different kinds of information have been considered: in [2], the different spectral shapes of each source are learned on isolated sounds and are then used to decompose the mixture. In [3], source signals are used as a side information in a coder/decoder scheme. In [4, 5], an aligned MIDI file is used to guide the separation of instruments in music pieces. In [6], textual information is used to perform separation.

In this paper we propose to use a speech signal imitating a target speech sound to guide the separation. More particularly, we use dialog dubs provided by the user as a reference signal to separate the dialog from the music and effects in a film or TV soundtrack.

A similar approach was already proposed in [7, 8] using a method based on probabilistic latent component analysis (PLCA) for isolating sounds in a mixture from the presentation of a humming query. This query mimics the desired target to be extracted and serves as a prior in the PLCA decomposition of the mixture. Unfortunately the methods suffers from robustness with respect to small time misalignements, pitch modifications and differences in equalization. We thus propose a new method that explicitly models these three issues. It is based on the adaptation of the power spectrogram of the provided speech guide both in pitch and time in a non-negative decomposition framework.

In section 2, we present the spectrogram model that we use for describing both the target signal to extract and the background. In section 3, we present the algorithm used to estimate the model parameters. In section 4, we provide experimental results that show that our algorithm outperforms state-of-the-art methods. Finally, conclusions are drawn in section 5.

**Notation**

- Matrices are denoted by bold capital letters: $\mathbf{M}$.
- Vectors are denoted by bold lower case letters: $\mathbf{v}$.
- Matrix or vector sizes are denoted by capital letters: $T$, whereas indexes are denoted with lower case letters: $t$.
- Scalars are denoted by italic lower case letters: $s$. The coefficients at row $f$ and column $t$ of matrix $\mathbf{M}$ is denoted by $m_{ft}$.

## 2. MODEL

As previously described, the goal of the proposed algorithm is to isolate a target sound, which is imitated by the user, from a background sound. As shown in figure 1, a typical scenario is a movie soundtrack where dialogs, effects and music are all mixed together and we want either to isolate the dialogs or to remove them: the user roughly dubs the dialogs and the dub signal is used as a guide. Another scenario would be a song in which we want to isolate or remove lead vocals: the user would then roughly re-sing it and this signal would be used as a guide. In both scenarios, there can be differences both in pitch and time between the guide and the target signals. This is why we propose a model which deals with these two issues. The guide signal is not restricted to be user-provided speech and could be anything else that is close enough to the target signal, such as a slightly modified version of the target signal (e.g., the target signal could have playback speed differences with the guide signal).

We work with power log-frequency spectrograms that are defined as the squared modulus of the constant-Q transform [9] (CQT) of the waveforms. The shift invariance property of the CQT (a pitch modification can be modeled by a simple vertical shift) is central in the model we propose.

For the sake of clarity, we will present the signal model for mono signals only although it can be easily generalized to multichannel signals using an extra panoramic parameter. In the experimental section, we actually use a stereo signal model.

This work was supported by the EUREKA Eurostars i3DMusic project funded by Oseo.

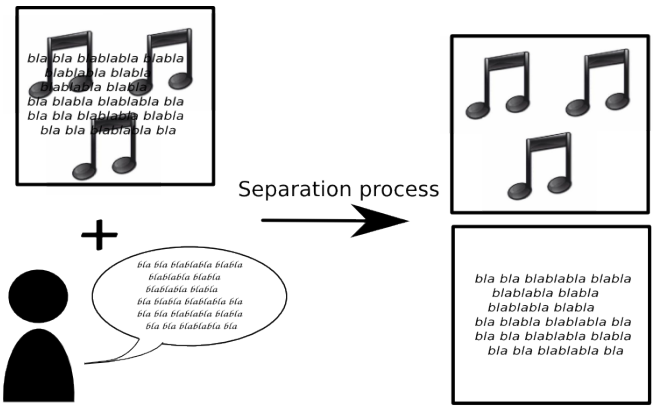


**Fig. 1**. Separation scenario.

The power spectrogram of the user guide sound is a $F \times T$ matrix denoted by $\mathbf{V}^g$ and the power spectrogram of the mix to be decomposed is a $F \times T$ matrix denoted by $\mathbf{V}$ (the guide and mix signals are supposed to be of the same temporal length, but it is straightforward to consider slightly different lengths). A common assumption in non-negative spectrogram decomposition is to model the spectrogram of the mix as the sum of the spectrogam of the target signal $\hat{\mathbf{V}}^t$ and the spectrogram of the background signal $\hat{\mathbf{V}}^b$ (the hat refers to the quantity to be estimated). We aim at estimating both these model spectrograms in order to approximate the mix spectrogram $\mathbf{V}$:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \hat{\mathbf{V}}^t + \hat{\mathbf{V}}^b \tag{1}$$

Once $\hat{\mathbf{V}}^t$ and $\hat{\mathbf{V}}^b$ are estimated, the separation is then performed with Wiener filtering using the estimated power spectrogram as a time-varying power spectral density.

### 2.1. Target signal model

The model of the target source is derived from the guide spectrogram $\mathbf{V}^g$ using three different kinds of adaptation:

- a pitch shift operator is applied in order to compensate pitch differences between guide and target signals.
- a synchronization matrix is used to prevent slight temporal misalignments between guide and target signals as proposed in [10].
- a global adaptation filter is also applied in order to correct global spectral differences (equalization) between guide and target signals.

All these parameters are constrained to be non-negative.

#### *2.1.1. Pitch shift operator*

The pitch shift operator is a $\Phi \times T$ matrix $\mathbf{P}$. It acts as a vertical shift operator on each time frame $t$ of the guide spectrogram $\mathbf{V}^g$, as in [11]. Spectrograms being computed with a CQT, a shift of a spectrogram frame corresponds to a pitch modification. The operation can be written:

$$\mathbf{V}^g_{\text{shifted}} = \left( \sum_{\phi} \overset{\downarrow\phi}{\mathbf{V}}{}^g \text{diag}(\mathbf{P}_{\phi,:}) \right) \tag{2}$$

where $\overset{\downarrow\phi}{\mathbf{V}}{}^g$ stands for $\mathbf{V}^g$ shifted downward by $\phi$ bins (*i.e.* $[\overset{\downarrow\phi}{\mathbf{V}}{}^g]_{f,t} = [\mathbf{V}^g]_{f-\phi,t}$), and $\text{diag}(\mathbf{P}_{\phi,:})$ is the diagonal matrix with the $\phi$-th row of $\mathbf{P}$ as diagonal.

The pitch shift operator is supposed to model possible differences between instantaneous pitch of the guide and the target signals. In practice, only one shift should be kept (and not a linear combination of all the shifts) so that the correct shift will be tracked (see section 3).

#### *2.1.2. Synchronization matrix*

The $T \times T$ synchronization matrix $\mathbf{S}$ allows a temporal alignment margin between the guide spectrogram and the target signal spectrogram: time frames of the target spectrogram are modeled as a linear combination of the neighboring time frames of the (pitch-shifted) guide spectrogram[1]. This makes it possible to take time misalignements (even time-varying ones) into account. This adaptation is expressed as:

$$\mathbf{V}^g_{\text{sync}} = \mathbf{V}^g_{\text{shifted}} \mathbf{S} \tag{3}$$

where $\mathbf{S}$ is a band matrix *i.e.* there exists $w \in \mathbb{N}$ such that for all $(t_1, t_2)$, if $|t_1 - t_2| > w$, then $s_{t_1 t_2} = 0$. The width $w$ of the central band corresponds to the misalignement tolerance in time frames. A large width will thus result in a large tolerance but at the price of a worse estimation of the model parameters. Only one (or a few) frames of the guide signal correspond to a frame of the guide signal and the correct synchronization may be tracked in the matrix $\mathbf{S}$.

#### *2.1.3. Adaptation filter*

As proposed in [12], the adaptation filter parameter is a $F \times 1$ vector $\mathbf{f}$ that acts as a global filter on the model. The global spectrogram model of the target signal is then:

$$\hat{\mathbf{V}}^t = \text{diag}\,(\mathbf{f}) \left( \sum_{\phi} \overset{\downarrow\phi}{\mathbf{V}}{}^g \text{diag}(\mathbf{P}_{\phi,:}) \right) \mathbf{S} \tag{4}$$

where $\text{diag}\,(\mathbf{f})$ is a diagonal matrix with $\mathbf{f}$ as the main diagonal.

### 2.2. Background signal model

As we do not have any information about the content of the background, we use a very common generic model. Standard NMF seems to be well adapted to this purpose. Thus, the power spectrogram of the background signal is modeled as:

$$\hat{\mathbf{V}}^b = \mathbf{W}\mathbf{H} \tag{5}$$

where $\mathbf{W}$ is a $F \times R$ non-negative matrix and $\mathbf{H}$ a $R \times T$ non-negative matrix with $R \ll F, T$ (the choice of $R$ is important and depends on the application). Columns of $\mathbf{W}$ can be thought of as atomic spectral templates and $\mathbf{H}$ as the activation weights of these templates over time.

## 3. ALGORITHM

### 3.1. First estimation of parameters

In order to estimate the parameters of the target signal model and the background signal model, an element-wise divergence cost function

[1]For simplicity, a square synchronization matrix is assumed, but an extension to non-square matrices is straightforward.

is minimized with respect to these parameters:

$$\mathcal{C}(\mathbf{\Theta}) = D(\mathbf{V}|\hat{\mathbf{V}}^t + \hat{\mathbf{V}}^b) = \sum_{f,t} d(v_{ft}|\hat{v}^t_{ft} + \hat{v}^b_{ft}) \quad (6)$$

In this paper, we use the Itakura-Saito divergence which is a very popular divergence in audio processing:

$$d(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \quad (7)$$

The minimization is done with multiplicative update rules which are successively applied to each of the model parameters: $\mathbf{W}$, $\mathbf{H}$, $\mathbf{f}$, $\mathbf{S}$ and $\mathbf{P}$.

We follow a classical approach to derive the update rules: the gradient of the cost function with respect to each parameter is written as a difference of two positive terms and the update rule is a multiplication by the ratio of these two terms. This notably ensures that parameters remain non-negative at each update and become constant if the partial derivative of the cost function with respect to the considered parameter tends to zero. Moreover, parameters evolve in a local descent direction.

We thus get the following update rules for the parameters of the target spectogram model:

$$\mathbf{f} \leftarrow \mathbf{f} \odot \frac{\left(\left(\sum_\phi \overset{\downarrow\phi}{\mathbf{V}^g}\text{diag}(\mathbf{P}_{\phi,:})\mathbf{S}\right) \odot \mathbf{V} \odot \hat{\mathbf{V}}^{\odot -2}\right)\mathbf{1}_T}{\left(\left(\sum_\phi \overset{\downarrow\phi}{\mathbf{V}^g}\text{diag}(\mathbf{P}_{\phi,:})\mathbf{S}\right) \odot \hat{\mathbf{V}}^{\odot -1}\right)\mathbf{1}_T} \quad (8)$$

$$\mathbf{S} \leftarrow \mathbf{S} \odot \frac{\left(\sum_\phi \text{diag}\left(\mathbf{f}\right) \overset{\downarrow\phi}{\mathbf{V}^g}\text{diag}(\mathbf{P}_{\phi,:})\right) \odot \mathbf{V} \odot \hat{\mathbf{V}}^{\odot -2}}{\left(\sum_\phi \text{diag}\left(\mathbf{f}\right) \overset{\downarrow\phi}{\mathbf{V}^g}\text{diag}(\mathbf{P}_{\phi,:})\right) \odot \hat{\mathbf{V}}^{\odot -1}} \quad (9)$$

$$\mathbf{P}_{\phi,:} \leftarrow \mathbf{P}_{\phi,:} \odot \frac{\mathbf{f}^T\left(\overset{\downarrow\phi}{\mathbf{V}^g} \odot \left(\left(\mathbf{V} \odot \hat{\mathbf{V}}^{\odot -2}\right)\mathbf{S}^T\right)\right)}{\mathbf{f}^T\left(\overset{\downarrow\phi}{\mathbf{V}^g} \odot \left(\hat{\mathbf{V}}^{\odot -1}\mathbf{S}^T\right)\right)} \quad (10)$$

where $\odot$ stands for the element-wise matrix (or vector) product, $(.)^{\odot(.)}$ stands for element-wise matrix exponentiation, $(.)^T$ stands for the matrix transposition, $\mathbf{1}_T$ is a $T \times 1$ vector with all coefficients equal to 1 and $\mathbf{P}_{\phi,:}$ is the $\phi$-th row of $\mathbf{P}$.

The update rules of $\mathbf{W}$ and $\mathbf{H}$ are standard multiplicative update rules for NMF with Itakura-Saito cost function [13].

All parameters are initialized with random non-negative values.

### 3.2. Shift value estimation and refining

As already stated, a target spectrogram frame is modeled (up to filter adaptation and synchronization) as a linear combination of pitch-shifted versions of the corresponding frame in the guide spectrogram. As our model intends to describe small differences in pitch, only one shift should be kept at each time frame. We thus introduce a tracking step to estimate the correct pitch shift value at each time frame within matrix $\mathbf{P}$. We propose to use Viterbi tracking such as in [14] where the tracking is done on a pitch matrix (estimation of the best pitch) instead of a pitch shift matrix.

Once the correct pitch shift has been tracked, coefficients of matrix $\mathbf{P}$ which do not correspond to the tracked path are set to 0. In practice we allow a small margin around the tracked pitch for two reasons: first, pitch shifts are quantized in the model but are continuous in real cases and second, the tracking algorithm might produce small errors. Then, parameters are reestimated using the update rules of section 3.1 with the thresholded version of $\mathbf{P}$ (as update rules are multiplicative, coefficients set to 0 will remain 0).

It should be noted that, as already suggested, it might also be necessary to track the right lag in the synchronization matrix using, for instance, dynamic time warping. We chose not to describe it here, as in the tested application the algorithm provided better results without synchronization tracking: this might be linked to a bad tracking or to the fact that thresholding the synchronization matrix might over-constrain the model.

## 4. EXPERIMENTAL RESULTS

### 4.1. Experimental setup

We tested the algorithm on a task of dialog isolation/removal in movie soundtracks. As a comparison, we tested three other separation algorithms: the first one is uninformed and is based on the main melody extraction method proposed in [14], which is a state-of-the-art blind method for this kind of task. The second one is the PLCA-based speech-informed separation algorithm proposed in [7]. The third one is the same algorithm as the first one but informed with instantaneous pitch of the target signal as in [15]. The comparison with the first algorithm intends to show that adding extra information can significantly improve separation results. The comparison with the second algorithm intends to show that with the same extra information, our algorithm performs better. The comparison with the third algorithm intends to show that our algorithm performs as well as another informed scenario but with a less tedious user input. We also present the results of a Wiener oracle for the CQT representation (oracles are obtained using the original sources to build CQT Wiener masks) which can be thought of as an upper performance bound.

We built a database of synthetic movie soundtracks. Each excerpt of the database was created mixing two different parts of a same movie soundtrack, the first one containing only dialog and the second one containing only music and effects. For each excerpt, mixing coefficients were estimated from parts of the same movie where dialog, music and effects were active all together and based on loudness[2] difference in order to create realistic mixes.

The database consists of 10 extracts from 5 different movies. All soundtracks were mixed down to mono signals. In all movies, dialogs are in english. All the excerpts were dubbed using the mix signal as a reference. All dubbings were done by the same male native english speaker. The same dubs were used for both speech informed algorithms (PLCA-based and the one that we propose).

Spectrograms were computed using the CQT implementation proposed in [9], with $f_{\text{min}} = 40$Hz, $f_{\text{max}} = 16000$Hz and 48 bins per octave.

### 4.2. Results

In order to quantify results we use standard metrics of source separation as proposed in [16]: Signal to Distorsion Ratio (SDR), Signal to Artefact Ratio (SAR) and Signal to Interference Ratio (SIR). Results are presented in figure 2 for the extracted dialog tracks and in 3 for the music and effects tracks. As can be seen, results of the blind method are significantly lower than any informed method in both cases, which confirms the benefits of informed methods. We can also notice that the PLCA-based speech-informed separation performs significantly worse than the method we propose. The comparison

[2]Loudness is defined following recommendation ITU-R BS.1770-2

between the pitch-informed method and ours is less clear: differences in terms of SDR are not significant (less than 0.2dB, which is about the same difference as between two consecutive runs with different initializations for all the proposed algorithms). Results in terms of SAR and SIR are about the opposite from the dialog extraction task to the dialog removal task. Thus, it is not possible to draw a conclusion from these metrics. As the differences are clearly audible (artefacts are quite different), we performed an internal blind listening test based on the MUSHRA protocol. 5 sound engineers were asked to rate the "usability" of each sound for the dialog extraction task only. A pairwise $t$-test on the listening test results showed that the results of our algorithm are globally preferred in terms of "usability" over the results of the pitch-informed algorithm ($p$-value $= 0.0017$). These last subjective results should be however taken with care since the number of participants was low and should thus be considered as a preliminary study.

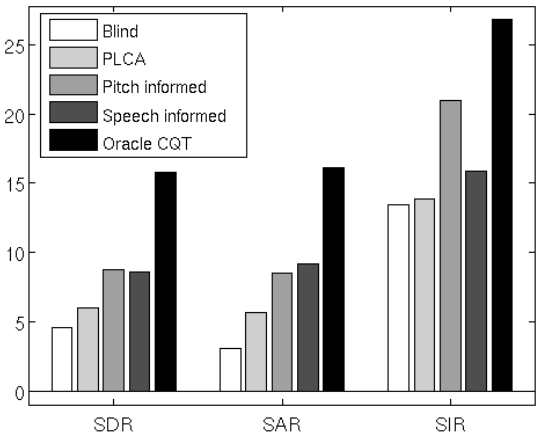


**Fig. 2**. Separation results for dialog extraction.

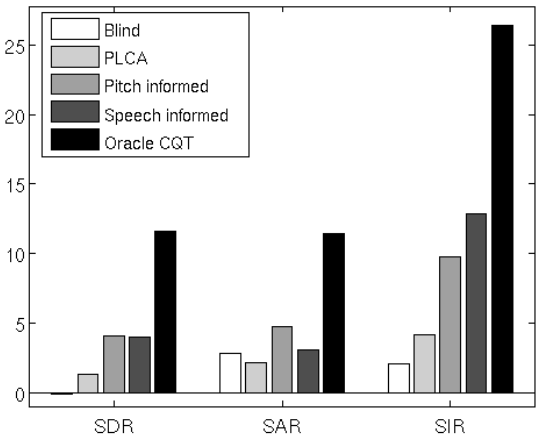


**Fig. 3**. Separation results for music and effects extraction (dialog removal).

## 5. CONCLUSION

In this paper we proposed a new method to perform source separation providing a spoken guide and showed that this method outperforms a state-of-the-art one. Future work should focus on speeding up the algorithm since the proposed algorithm is a bit slower than the ones we compared it with. Moreover, other kind of adaptation such as formant adaptation might also be considered in order to get a better fit of the target from the guide. Hard introduction of the guide spectrogram in the model of the target spectrogram could also be replaced with soft introduction using priors. Finally, as the system we propose does not assume any voice model on the guide signal, other kind of signal can be used as a guide, opening the way for other applications.

## 6. REFERENCES


[1] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[2] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *7th International Conference on Independent Component Analysis and Signal Separation*, London, UK, September 2007.

[3] Mathieu Parvaix, Laurent Girin, and Jean-Marc 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, August 2010.

[4] Romain Hennequin, Bertrand David, and Roland Badeau, "Score informed audio source separation using a parametric model of nonnegative spectrogram," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[5] Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel, "Evaluation of a score-informed source separation system," in *International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, August 2010.

[6] Luc Le Magoarou, Alexey Ozerov, and Ngoc Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial cofactorization," in *IEEE International Workshop on Machine Learning for Signal Processing*, Southapmton,UK, September 2013.

[7] Paris Smaragdis and Gautham J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 2009, pp. 69 – 72.

[8] Derry FitzGerald, "User assisted separation using tensor factorisations," in *European Signal Processing Conference*, Bucharest, Romania, August 2012, pp. 2412 – 2416.

[9] Christian Schrkhuber and Anssi Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain, July 2010.

[10] Agnes Pedone, Juan-Jose Burred, Simon Maller, and Pierre Leveau, "Phoneme-level text to audio synchronization on speech signals with background music," in *Annual Conference of the International Speech Communication Association*, Florence, Italy, September 2011.

[11] Mikkel N. Schmidt and Morten Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Paris, France, April 2006, vol. 3889 of *Lecture Notes in Computer Science (LNCS)*, pp. 700–707, Springer.

[12] Pierre Leveau, Simon Maller, Juan José Burred, and Xabier Jaureguiberry, "Convolutive common audio signal extraction," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 2011, pp. 165–168.

[13] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 11, no. 3, pp. 793–830, March 2009.

[14] Jean-Louis Durrieu, Gaël Richard, and Bertrand David, "An iterative approach to monaural musical mixture de-soloing," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 105 – 108.

[15] Jean-Louis Durrieu and Jean-Philippe Thiran, "Musical audio source separation based on user-selected f0 track," in *International Conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, March 2012, pp. 438–445.

[16] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.