LONG-TERM REVERBERATION MODELING FOR UNDER-DETERMINED AUDIO SOURCE SEPARATION WITH APPLICATION TO VOCAL MELODY EXTRACTION

Romain Hennequin
Deezer R&D
10 rue d'Athènes, 75009 Paris, France
rhennequin@deezer.com

François Rigaud
Audionamix R&D
171 quai de Valmy, 75010 Paris, France
francois.rigaud@audionamix.com

ABSTRACT

In this paper, we present a way to model long-term reverberation effects in under-determined source separation algorithms based on a non-negative decomposition framework. A general model for the sources affected by reverberation is introduced and update rules for the estimation of the parameters are presented. Combined with a well-known source-filter model for singing voice, an application to the extraction of reverberated vocal tracks from polyphonic music signals is proposed. Finally, an objective evaluation of this application is described. Performance improvements are obtained compared to the same model without reverberation modeling, in particular by significantly reducing the amount of interference between sources.

1. INTRODUCTION

Under-determined audio source separation has been a key topic in audio signal processing for the last two decades. It consists in isolating different meaningful 'parts' of the sound, such as for instance the lead vocal from the accompaniment in a song, or the dialog from the background music and effects in a movie soundtrack. Non-negative decompositions such as Non-negative Matrix Factorization [5] and its derivative have been very popular in this research area for the last decade and have achieved state-of-the art performances [3, 9, 12].

In music recordings, the vocal track generally contains reverberation that is either naturally present due to the recording environment or artificially added during the mixing process. For source separation algorithms, the effects of reverberation are usually not explicitly modeled and thus not properly extracted with the corresponding sources. Some studies [1, 2] introduce a model for the effect of spatial diffusion caused by the reverberation for a multichannel source separation application. In [7] a model for the dereverberation of spectrograms is presented for the case of long reverberations, *i.e.* when the reverberation time is longer than the length of the analysis window.

We propose in this paper to extend the model of reverberation proposed in [7] to a source separation application that allows extracting the reverberation of a specific source together with its dry signal. The reverberation model is introduced first in a general framework for which no assumption is made about the spectrogram of the dry sources. At this state, and as often in demixing application, the estimation problem is ill-posed (optimization of a non-convex cost function with local minima, result highly dependent on the initialization, ...) and requires the incorporation of some knowledge about the source signals. In [7], this issue is dealt with a sparsity prior on the unreverberated spectrogram model. Alternatively, the spectrogram sources can be structured by using models of non-negative decompositions with constraints (*e.g.* harmonic structure of the source's tones, sparsity of the activations) and/or by guiding the estimation process with prior information (*e.g.* source activation, multi-pitch transcription). Thus in this paper we propose to combine the generic reverberation model with a well-known source/filter model of singing voice [3]. A modified version of the original voice extraction algorithm is described and evaluated on an application to the extraction of reverberated vocal melodies from polyphonic music signals.

Note that unlike usual application of reverberation modeling, we do not aim at extracting dereverbated sources but we try to extract accurately both the dry signal and the reverberation within the same track. Thus, the designation source separation is not completely in accordance with our application which targets more precisely *stem* separation.

The rest of the paper is organized as follows: Section 2 presents the general model for a reverberated source and Section 3 introduces the update rule used for its estimation. In Section 4, a practical implementation for which the reverberation model is combined with a source/filter model is presented. Then, Section 5 presents experimental results that demonstrate the ability of our algorithm to extract properly vocals affected by reverberation. Finally, conclusions are drawn in Section 6.

Attribution: Romain Hennequin, François Rigaud. "Long-term reverberation modeling for under-determined audio source separation with application to vocal melody extraction", 17th International Society for Music Information Retrieval Conference, 2016. This work has been done when the first author was working for Audionamix.

Notation

- Matrices are denoted by bold capital letters: $\mathbf{M}$. The coefficients at row f and column t of matrix $\mathbf{M}$ is denoted by $\mathbf{M}_{f,t}$.
- 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.
- $\odot$ stands for element-wise matrix multiplication (Hadamard product) and $\mathbf{M}^{\odot\lambda}$ stands for element-wise exponentiation of matrix $\mathbf{M}$ with exponent λ.

2. GENERAL MODEL

For the sake of clarity, we will present the signal model for mono signals only although it can be easily generalized to multichannel signals as in [6]. In the experimental Section 5, a stereo signal model is actually used.

2.1 Non-negative Decomposition

Most source separation algorithms based on a non-negative decomposition assume that the non-negative mixture spectrogram $\mathbf{V}$ (usually the modulus or the squared modulus of a time-frequency representation such as the Short Time Fourier Transform (STFT)) which is a $F \times T$ non-negative matrix, can be approximated as the sum of K source model spectrograms $\hat{\mathbf{V}}^k$, which are also non-negative:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{k=1}^{K} \hat{\mathbf{V}}^k \tag{1}$$

Various structured matrix decomposition have been proposed for the source models $\hat{\mathbf{V}}^k$, such as, to name a few, standard NMF [8], source/filter modeling [3] or harmonic source modeling [10].

2.2 Reverberation Model

In the time-domain, a time-invariant reverberation can be accurately modeled using a convolution with a filter and thus be written as:

$$\mathbf{y} = \mathbf{h} * \mathbf{x}, \tag{2}$$

where $\mathbf{x}$ is the dry signal, $\mathbf{h}$ is the impulse response of the reverberation filter and $\mathbf{y}$ is the reverberated signal.

For short-term convolution, this expression can be approximated by a multiplication in the frequency domain such as proposed in [6] :

$$\overline{\mathbf{y}}_t = \overline{\mathbf{h}} \odot \overline{\mathbf{x}}_t, \tag{3}$$

where $\overline{\mathbf{x}}_t$ (respectively $\overline{\mathbf{y}}_t$) is the modulus of the t-th frame of the STFT of $\mathbf{x}$ (respectively $\mathbf{y}$) and $\overline{\mathbf{h}}$ is the modulus of the Fourier transform of $\mathbf{h}$.

For long-term convolution, this approximation does not hold. The support of typical reverberation filters are generally greater than half a second which is way too long for a STFT analysis window in this kind of application. In this case, as suggested in [7], we can use an other approximation which is a convolution in each frequency channel :

$$\overline{\mathbf{y}}^f = \overline{\mathbf{h}}^f * \overline{\mathbf{x}}^f, \tag{4}$$

Where $\overline{\mathbf{y}}^f$, $\overline{\mathbf{h}}^f$ and $\overline{\mathbf{x}}^f$ are the f-th frequency channel of the STFT of respectively $\mathbf{y}$, $\mathbf{h}$ and $\mathbf{x}$.

Then, starting from a dry spectrogram model $\hat{\mathbf{V}}^{\text{dry},k}$ of a source with index k, the reverberated model of the same source is obtained using the following non-negative approximation:

$$\hat{\mathbf{V}}^{\text{rev},k}_{f,t} = \sum_{\tau=1}^{T_k} \hat{\mathbf{V}}^{\text{dry},k}_{f,t-\tau+1} \mathbf{R}^k_{f,\tau} \tag{5}$$

where $\mathbf{R}^k$ is the $F \times T_k$ non-negative reverberation matrix of model k to be estimated.

The model of Equation (5) makes it possible to take long-term effects of reverberation into account and generalizes short-term convolution models as proposed in [6] since when $T_k = 1$, the model corresponds to the short-term convolution approximation.

3. ALGORITHM

3.1 Non-negative decomposition algorithms

The approximation of Equation (1) is generally quantified using a divergence (a measure of dissimilarity) between $\mathbf{V}$ and $\hat{\mathbf{V}}$ to be minimized with respect to the set of parameters $\mathbf{\Lambda}$ of all the models:

$$\mathcal{C}(\mathbf{\Lambda}) = D(\mathbf{V}|\hat{\mathbf{V}}(\mathbf{\Lambda})) \tag{6}$$

A commonly used class of divergence is the element-wise β-divergence which encompasses the Itakura-Saito divergence ($\beta = 0$), the Kullback-Leibler divergence ($\beta = 1$) and the squared Frobenius distance ($\beta = 2$) [4]. The global cost then writes:

$$\mathcal{C}(\mathbf{\Lambda}) = \sum_{f,t} d_\beta(\mathbf{V}_{f,t}|\hat{\mathbf{V}}_{f,t}(\mathbf{\Lambda})). \tag{7}$$

The problem being not convex, the minimization is generally done using alternating update rules on each parameters of Λ. The update rule for a parameter $\mathbf{\Theta}$ is commonly obtained using an heuristic consisting in decomposing the gradient of the cost-function with respect to this parameter as a difference of two positive terms, such as

$$\nabla_{\mathbf{\Theta}}\mathcal{C} = P_{\mathbf{\Theta}} - M_{\mathbf{\Theta}}, \quad P_{\mathbf{\Theta}} \geq 0,\ M_{\mathbf{\Theta}} \geq 0, \tag{8}$$

and then by updating the parameter according to:

$$\mathbf{\Theta} \leftarrow \mathbf{\Theta} \odot \frac{M_{\mathbf{\Theta}}}{P_{\mathbf{\Theta}}}. \tag{9}$$

This kind of update rule ensures that the parameter remains non-negative. Moreover the parameter is updated in a direction descent or remains constant if the partial derivative is zero. In some cases (including the update rules

we will present), it is possible to prove using a Majorize-Minimization (MM) approach [4] that the multiplicative update rules actually lead to a decrease of the cost function.

Using such an approach, the update rules for a standard NMF model $\hat{\mathbf{V}} = \mathbf{WH}$ can be expressed as:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left(\hat{\mathbf{V}}^{\odot\beta-2} \odot \mathbf{V}\right)}{\mathbf{W}^T \hat{\mathbf{V}}^{\odot\beta-1}}, \quad (10)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left(\hat{\mathbf{V}}^{\odot\beta-2} \odot \mathbf{V}\right) \mathbf{H}^T}{\hat{\mathbf{V}}^{\odot\beta-1} \mathbf{H}^T}. \quad (11)$$

3.2 Estimation of the reverberation matrix

When dry models $\hat{\mathbf{V}}^{\text{dry},k}$ are fixed, the reverberation matrix can be estimated using the following update rule applied successively on each reverberation matrix:

$$\mathbf{R}^k \leftarrow \mathbf{R}^k \odot \frac{\left(\hat{\mathbf{V}}^{\odot\beta-2} \odot \mathbf{V}\right) *_t \hat{\mathbf{V}}^{\text{dry},k}}{\hat{\mathbf{V}}^{\odot\beta-1} *_t \hat{\mathbf{V}}^{\text{dry},k}} \quad (12)$$

where $*_t$ stands for time-convolution:

$$\left[\hat{\mathbf{V}}^{\odot\beta-1} *_t \hat{\mathbf{V}}^{\text{dry},k}\right]_{f,\tau} = \sum_{\tau=t}^{T} \hat{\mathbf{V}}_{f,\tau}^{\odot\beta-1} \hat{\mathbf{V}}_{f,\tau-t+1}^{\text{dry},k}. \quad (13)$$

The update rule (12) obtained using the procedure described in Section 3.1 ensures the non-negativity of $\mathbf{R}^k$. This update can be obtained using the MM approach which ensures that the cost-function will not increase.

3.3 Estimation with other free model

A general drawback of source separation models that do not explicitly account for reverberation effects is that the reverberation affecting a given source is usually spread among all separated sources. However, this issue can still arise with a proper model of reverberation if the dry model of the reverberated source is not constrained enough. Indeed using the generic reverberation model of Equation (5), the reverberation of a source can still be incorrectly modeled during the optimization by other source models having more degrees of freedom. A possible solution for enforcing a correct optimization of the reverberated model is to further constrain the structure of the dry model spectrogram, *e.g.* through the inclusion of sparsity or pitch activation constraints, and potentially to adopt a sequential estimation scheme. For instance, first discarding the reverberation model, a first rough estimate of the dry source may be produced. Second, considering the reverberation model, the dry model previously estimated can be refined while estimating at the same time the reverberation matrix. Such an approach is described in the following section for a practical implementation of the algorithm to the problem of lead vocal extraction from polyphonic music signals.

4. APPLICATION TO VOICE EXTRACTION

In this section we propose an implementation of our reverberation model in a practical case: we use Durrieu's algorithm [3] for lead vocal isolation and add the reverberation model over the voice model.

4.1 Base voice extraction algorithm

Durrieu's algorithm for lead vocal isolation in a song is based on a source/filter model for the voice.

4.1.1 Model

The non-negative mixture spectrogram model consists in the sum of a voice spectrogram model based on a source/filter model and a music spectrogram model based on a standard NMF:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \hat{\mathbf{V}}^{\text{voice}} + \hat{\mathbf{V}}^{\text{music}}. \quad (14)$$

The voice model is based on a source/filter speech production model:

$$\hat{\mathbf{V}}^{\text{voice}} = (\mathbf{W}_{\text{F0}}\mathbf{H}_{\text{F0}}) \odot (\mathbf{W}_{\text{K}}\mathbf{H}_{\text{K}}). \quad (15)$$

The first factor $(\mathbf{W}_{\text{F0}}\mathbf{H}_{F0})$ is the source part corresponding to the excitation of the vocal folds: $\mathbf{W}_{\text{F0}}$ is a matrix of fixed harmonic atoms and $\mathbf{H}_{\text{F0}}$ is the activation of these atoms over time. The second factor $(\mathbf{W}_{\text{K}}\mathbf{H}_{\text{K}})$ is the filter part corresponding to the resonance of the vocal tract: $\mathbf{W}_{\text{K}}$ is a matrix of smooth filter atoms and $\mathbf{H}_{\text{K}}$ is the activation of these atoms over time.

The background music model is a generic NMF:

$$\hat{\mathbf{V}}^{\text{music}} = \mathbf{W}_{\text{R}}\mathbf{H}_{\mathbf{R}}. \quad (16)$$

4.1.2 Algorithm

Matrices $\mathbf{H}_{\text{F0}}$, $\mathbf{W}_{\text{K}}$, $\mathbf{H}_{\text{K}}$, $\mathbf{W}_{\text{R}}$ and $\mathbf{H}_{\text{R}}$ are estimated minimizing the element-wise Itakura-Saito divergence between the original mixture power spectrogram and the mixture model:

$$\mathcal{C}(\mathbf{H}_{\text{F0}}, \mathbf{W}_{\text{K}}, \mathbf{H}_{\text{K}}, \mathbf{W}_{\text{R}}, \mathbf{H}_{\text{R}}) = \sum_{f,t} d_{\text{IS}}(\mathbf{V}_{f,t}|\hat{\mathbf{V}}_{f,t}), \quad (17)$$

where $d_{\text{IS}}(x, y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$. The minimization is achieved using multiplicative update rules.

The estimation is done in three steps:

1. A first step of parameter estimation is done using iteratively the multiplicative update rules.

2. The matrix $\mathbf{H}_{\text{F0}}$ is processed using a Viterbi decoding for tracking the main melody and is then thresholded so that coefficients too far from the melody are set to zero.

3. Parameters are re-estimated as in the first step but using the thresholded version of $\mathbf{H}_{\text{F0}}$ for the initialization.

4.2 Inclusion of the reverberation model

As stated in Section 3, the dry spectrogram model (*i.e.* the spectrogram model for the source without reverberation) has to be sufficiently constrained in order to accurately estimate the reverberation part. This constraint is here obtained through the use of a fixed harmonic dictionary $\mathbf{W}_{\text{F0}}$ and mostly, by the thresholding of the matrix $\mathbf{H}_{\text{F0}}$ that enforces the sparsity of the activations.

We thus introduce the reverberation model after the step of thresholding of the matrix $\mathbf{H}_{\text{F0}}$. The two first steps then remains the same as presented in Section 4.1.2. In the third step, the dry voice model of Equation (15) is replaced by a reverberated voice model following Equation (5):

$$\hat{\mathbf{V}}_{f,t}^{\text{rev. voice}} = \sum_{\tau=1}^{T} \hat{\mathbf{V}}_{f,t-\tau+1}^{\text{voice}} \mathbf{R}_{f,\tau}. \tag{18}$$

For the parameter re-estimation of step 3, the multiplicative update rule of $\mathbf{R}$ is given by Equation (12). For the other parameters of the voice model, the update rules from [3] are modified to take the reverberation model into account:

$$\mathbf{H}_{\text{F0}} \leftarrow \mathbf{H}_{\text{F0}} \odot \frac{\mathbf{W}_{\text{F0}}^T \left((\mathbf{W}_{\text{K}}\mathbf{H}_{\text{K}}) \odot \left(\mathbf{R} *_t (\hat{\mathbf{V}}^{\odot\beta-2} \odot \mathbf{V}) \right) \right)}{\mathbf{W}_{\text{F0}}^T \left((\mathbf{W}_{\text{K}}\mathbf{H}_{\text{K}}) \odot \left(\mathbf{R} *_t \hat{\mathbf{V}}^{\odot\beta-1} \right) \right)} \tag{19}$$

$$\mathbf{H}_{\text{K}} \leftarrow \mathbf{H}_{\text{K}} \odot \frac{\mathbf{W}_{\text{K}}^T \left((\mathbf{W}_{\text{F0}}\mathbf{H}_{\text{F0}}) \odot \left(\mathbf{R} *_t (\hat{\mathbf{V}}^{\odot\beta-2} \odot \mathbf{V}) \right) \right)}{\mathbf{W}_{\text{K}}^T \left((\mathbf{W}_{\text{F0}}\mathbf{H}_{\text{F0}}) \odot \left(\mathbf{R} *_t \hat{\mathbf{V}}^{\odot\beta-1} \right) \right)} \tag{20}$$

$$\mathbf{W}_{\text{K}} \leftarrow \mathbf{W}_{\text{K}} \odot \frac{\left((\mathbf{W}_{\text{F0}}\mathbf{H}_{\text{F0}}) \odot \left(\mathbf{R} *_t (\hat{\mathbf{V}}^{\odot\beta-2} \odot \mathbf{V}) \right) \right) \mathbf{H}_{\text{K}}^T}{\left((\mathbf{W}_{\text{F0}}\mathbf{H}_{\text{F0}}) \odot \left(\mathbf{R} *_t \hat{\mathbf{V}}^{\odot\beta-1} \right) \right) \mathbf{H}_{\text{K}}^T} \tag{21}$$

The update rules for the parameters of the music model ($\mathbf{H}_{\text{R}}$ and $\mathbf{W}_{\text{R}}$), are unchanged and thus identical to those given in Equations (10) and (11).

5. EXPERIMENTAL RESULTS

5.1 Experimental setup

We tested the reverberation model that we proposed with the algorithm presented in Section 4 on a task of lead vocal extraction in a song. In order to assess the improvement of our model over the existing one, we ran the separation with and without reverberation modeling.

We used a database composed of 9 song excerpts of professionally produced music. The total duration of all excerpts was about 10 minutes. As the use of reverberation modeling only makes sense if there is a significant amount of it, all the selected excerpts contains a fair amount of reverberation. This reverberation was already present in the separated tracks and was not added artificially by ourselves. On some excerpts, the reverberation is time-variant: active on some parts and inactive on other, ducking echo effect ... Some short excerpts, as well as the separation results, can be played on the companion website [1] .

Spectrograms were computed as the squared modulus of the STFT of the signal sampled at 44100 Hz, with 4096-sample (92.9 ms) long Hamming window with 75% overlap. The length T of the reverberation matrix was arbitrarily fixed to 52 frames (which corresponds to about 1.2 s) in order to be sufficient for long reverberations.

5.2 Results

In order to quantify the results we use standard metrics of source separation as described in [11]: Signal to Distorsion Ratio (SDR), Signal to Artefact Ratio (SAR) and Signal to Interference Ratio (SIR).

The results are presented in Figure 1 for the evaluation of the extracted voice signals and in Figure 2 for the extracted music signals. The oracle performance, obtained using the actual spectrograms of the sources to compute the separation masks, are also reported. As we can see, adding the reverberation modeling increases all these metrics. The SIR is particularly increased in Figure 1 (more than 5dB): this is mainly because without the reverberation model, a large part of the reverberation of the voice leaks in the music model. This is a phenomenon which is also clearly audible in excerpts with strong reverberation: using the reverberation model, the long reverberation tail is mainly heard within the separated voice and is almost not audible within the separated music. In return, extracted vocals with the reverberation model tend to have more audible interferences. This result is in part due to the fact that the pre-estimation of the dry model (step 1 and 2 of the base algorithm) is not interference-free, so that applying the reverberation model increases the energy of these interferences.

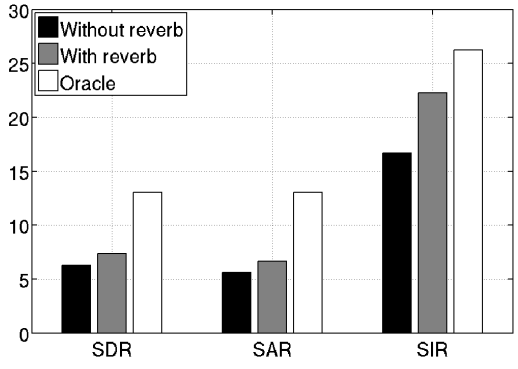

Figure 1. Experimental separation results for the voice stem.

[1] http://romain-hennequin.fr/En/demo/reverb_separation/reverb.html

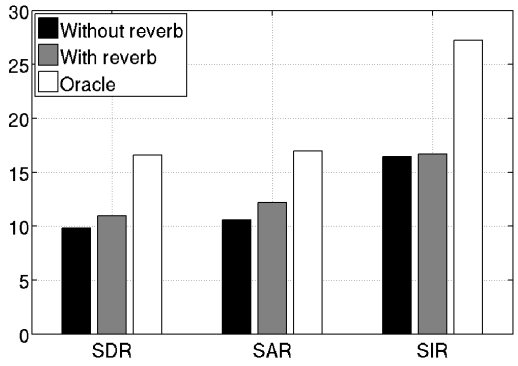

Figure 2. Experimental separation results for the music stem.

6. CONCLUSION

In this paper we proposed a method to model long-term effects of reverberation in a source separation application for which a constrained model of the dry source is available. Future work should focus on speeding up the algorithm since, multiple convolutions at each iteration can be time-consuming. Developing methods to estimate the reverberation duration (of a specific source within a mix) would also make it possible to automate the whole process. It could also be interesting to add spatial modeling for multichannel processing using full rank spatial variance matrix and multichannel reverberation matrices.

7. REFERENCES

[1] Simon Arberet, Alexey Ozerov, Ngoc Q. K. Duong, Emmanuel Vincent, Frédéric Bimbot, and Pierre Vandergheynst. Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In *International Conference on Information Sciences, Signal Processing and their applications*, pages 1–4, May 2010.

[2] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830 – 1840, Sept 2010.

[3] 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)*, pages 105–108, Taipei, Taiwan, April 2009.

[4] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, September 2011.

[5] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[6] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550–563, 2010.

[7] Rita Singh, Bhiksha Raj, and Paris Smaragdis. Latent-variable decomposition based dereverberation of monaural and multi-channel signals. In *IEEE International Conference on Audio and Speech Signal Processing*, Dallas, Texas, USA, March 2010.

[8] Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177 – 180, New Paltz, NY, USA, October 2003.

[9] 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.

[10] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):528–537, March 2010.

[11] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, July 2006.

[12] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, March 2007.