Informed Audio Source Separation: A Comparative Study

Antoine Liutkus, Stanislaw Gorlow, Nicolas Sturmel, Shuhua Zhang, Laurent Girin, Roland Badeau, Laurent Daudet, Sylvain Marchand, Gael Richard

INFORMED AUDIO SOURCE SEPARATION: A COMPARATIVE STUDY

Antoine Liutkus[1] *Stanislaw Gorlow*[2] *Nicolas Sturmel*[3] *Shuhua Zhang*[4]
Laurent Girin[4] *Roland Badeau*[1] *Laurent Daudet*[3] *Sylvain Marchand*[5] *Gaël Richard*[1]

[1] Institut Telecom, Telecom ParisTech, CNRS LTCI, France
[2] LaBRI, CNRS, Univ. Bordeaux 1, Talence, France
[3]Institut Langevin, CNRS, ESPCI-ParisTech, Univ. Paris Diderot, Paris, France
[4]GIPSA-Lab, Grenoble-INP, Grenoble, France
[5]Lab-STICC, CNRS, Univ. Western Brittany, Brest, France

ABSTRACT

The goal of source separation algorithms is to recover the constituent sources, or audio objects, from their mixture. However, blind algorithms still do not yield estimates of sufficient quality for many practical uses. Informed Source Separation (ISS) is a solution to make separation robust when the audio objects are known during a so-called encoding stage. During that stage, a small amount of side information is computed and transmitted with the mixture. At a decoding stage, when the sources are no longer available, the mixture is processed using the side information to recover the audio objects, thus greatly improving the quality of the estimates at a cost of additional bitrate which depends on the size of the side information. In this study, we compare six methods from the state of the art in terms of quality versus bitrate, and show that a good separation performance can be attained at competitive bitrates.

1. INTRODUCTION

Active listening, which is the possibility to independently manipulate, mute or equalize audio objects within their mixture [2, 10] has recently aroused much interest. In the context of music signal processing, an audio object can be identified with a sound *source* such as an instrument or a singing voice. All audio objects are mixed together into a mixture, which is possibly multichannel as in the common stereophonic case. From a practical point of view, active listening scenarios require the availability not only of the mixture, but also of the audio objects. Whereas some studies concentrate on a *coding* strategy [5, 11] to convey the audio objects, others focus on a *source separation* approach [13, 12, 9, 6]. The common idea of all these techniques is to make use of the mixture in order to reduce the required bitrate to transmit the audio objects. Indeed, a straightforward solution would be to separately encode the audio objects using audio coders. This solution is however not optimal in terms of bitrate because it amounts to encode several times the same information: once within the mixture and once alone for the transmission of the audio objects. Indeed, transmission of the mixture is considered mandatory in most cases.

Source separation consists in separating different signals from their mixture. Unfortunately, the performance of blind source separation, i.e. given the mixtures only, is still very dependent on the signals considered and does not systematically provide separated signals of sufficient quality for active listening applications. In *informed source separation* (ISS), source separation is improved by transmitting to the separation module some additional side information that strongly enhances the quality of the estimates. The general block diagram of ISS is presented in Fig. 1. Given the (mono) audio objects and the multichannel mixture, a small side information Θ is computed in the encoding stage and transmitted with the mixture. In the decoding stage, it allows to recover the objects given the mixture only. Many strategies can fit in this general framework, which are mainly distinguished by the assumptions made on the signals and by the corresponding separation algorithms. Indeed, it can be seen from Fig. 1 that the general ISS structure is independent from the separation method used at the decoder. Several approaches can hence be found in the literature [12, 11, 9, 6] and their common objective is to maximize the perceptual quality of the estimates while minimizing the size of the side information. Typical bitrates of the methods under study lie around 5-15kbps per object, which is very competitive compared to bitrates achieved by usual perceptual coders.

The purpose of this study is to present the general ideas of informed source separation and to briefly present four different techniques from the state of the art as well as to give the corresponding references. Then, all those techniques are evaluated on the same corpus in an extensive comparative evaluation. It is structured as follows. First, we present the framework considered and some notations in Section 2. Then, an overview of the different techniques under study is presented

This work is partly funded by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03)

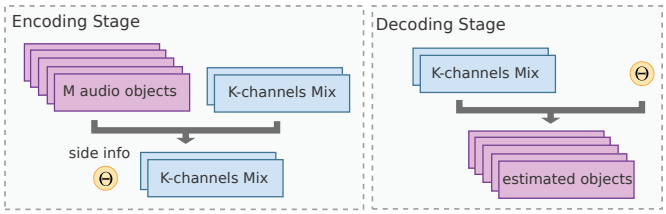

Fig. 1. High level block diagram of ISS. During the encoding stage, both the objects and the mixture are known and a small side information is computed. At the decoding stage, this information is used to perform robust source separation

in Section 3. Finally, an experimental comparison of all these techniques is found in Section 4, which allows to grasp some of the specificities of each method.

2. FRAMEWORK AND NOTATION

In the following, we assume that the M *audio objects* are defined as M regularly sampled times series s_m of the same length N_n. Furthermore, we suppose that a *mixing process* produces a K-channel mixture $\{y_k\}_{k=1\cdots K}$ from the audio objects. The Short-Term Fourier Transform (STFT) is written using capital letters, e.g. $S_m(\omega,t)$ and its squared modulus using bold capital letters, e.g. $\mathbf{S}_m(\omega,t)$. (ω,t) is called a Time-Frequency (TF) bin, where ω and t respectively stand for the frequency and frame index. N_ω and N_t are, respectively, the number of frequency bins and the number of frames of each transformed signal.

We consider linear and time-invariant mixing processes, since they are handled by all the techniques we review for ISS, contrarily to more complex models such as non-punctual or non-linear mixing [4, 15]. Thus, each audio object m is supposed to be mixed into each channel k through the use of a constant mixing filter a_{km}. When a_{km} reduces to a single gain coefficient, the mixing is called instantaneous. We assume that a_{km} are sufficiently short compared to the length of each frame so that the mixing process can be approximated in the STFT domain as:

$$Y(\omega,t) \approx A(\omega)\, S(\omega,t) \tag{1}$$

where

$$\begin{aligned} S(\omega,t) &= [S_1(\omega,t)\cdots S_M(\omega,t)]^\top \\ Y(\omega,t) &= [Y_1(\omega,t)\cdots Y_K(\omega,t)]^\top \end{aligned}$$

are $M\times 1$ and $K\times 1$ vectors gathering respectively all audio objects and all channels of the mixture at TF bin (ω,t). $A_{km}(\omega)$ is the frequency response of filter a_{km} at frequency bin ω. The $K\times M$ matrix $A(\omega)$ is called the *mixing matrix*. The objective of ISS is to compute some side information Θ that allows to recover good estimates $\hat{s}_m$ of the audio objects given the mixture $\{y_k\}_{k=1\cdots K}$. For the computation of Θ, we assume that y_k, s_m and A are all available.

3. FOUR SEPARATION METHODS

In this section, we briefly present the four different techniques for ISS that we compare in the following. Because of the space limitation, we cannot give the details concerning the models, which can be found in the corresponding references.

3.1. Local inversion

The first technique considered [12] is called "Local Inversion" in the following. It relies on the assumption that most audio signals can be considered *sparse* in the STFT domain, i.e. that only a small amount of TF bins have a significant magnitude, all others being close to zero. Consequently, for a given TF bin, it is much likely that only a subset of the objects have a noticeable energy. If we assume that at most K sources are active at some TF bin (ω,t), (1) becomes invertible locally for those sources, all others being set to zero.

For each TF bin, the technique then amounts to look for the combination of active objects that minimizes the squared error between $S(\omega,t)$ and its reconstruction, among the $\frac{M!}{K!(M-K)!}$ possible combinations. The side information Θ is thus the combination of active objects for each TF bin. The computational complexity of this method is $\mathcal{O}(N_\omega N_t M)$. A further refinement on [12] which permits variable bitrates is to send the combination of active objects only for TF bins whose magnitude lies above a perceptual threshold.

3.2. MMSE of locally stationary Gaussian processes

A second technique is based on the work in [9]. The sources are assumed to be locally stationary Gaussian processes. In that case, the TF coefficients of the mixtures can also be considered as Gaussian and the optimal MMSE estimator of the source signals is provided by the mean of the posterior (Gaussian) distribution of the sources given the mixtures [4], which writes:

$$\begin{aligned} &\mu_{\text{post}}(\omega,t) = \\ &\left(\text{diag}\mathbf{S}(\omega,t)\right) A(\omega)^H \left(A(\omega)\,\text{diag}\mathbf{S}(\omega,t)\,A(\omega)^H\right)^{-1} Y(\omega,t)\,. \end{aligned} \tag{2}$$

The side information considered by this method are A and the spectrograms $\mathbf{S}_m$ of the sources. [9] proposes to compress $\mathbf{S}_m$ using either image compression, e.g. JPEG, or Nonnegative Matrix Factorization (NMF). The corresponding techniques are denoted "MMSE-NMF" and "MMSE-JPG" respectively in the following. If the original $\mathbf{S}_m$ are used, the technique is called "Oracle MMSE". The complexity of MMSE techniques is $\mathcal{O}\left(N_\omega N_t K^3\right)$.

3.3. Iterative beamforming

The third technique, abbreviated as MVDR+PP, is the TF selective iterative spatial filtering algorithm presented in

[6]. It uses the minimum-variance distortionless response (MVDR) beamformer [3] in the filtering stage and during a post-processing stage, it readjusts the magnitudes of the estimated signal spectra in order to comply with the imposed equal-loudness constraint. The following modifications were made with regard to [6]. First, the spatial covariance matrices were approximated directly from the transmitted spectral envelopes and the mixing coefficients according to

$$\tilde{C}_{XX}(\omega,t) = A(\omega)\,\mathrm{diag}\hat{\mathbf{S}}(\omega,t)\,A(\omega)^H,$$

and no longer from the clustered data. Second, pairwise extraction of the two most dominant signal components was abandoned in favor of a one-by-one deflation strategy. In addition, the side information was compressed with bzip2[1], which uses Burrows-Wheeler block sorting text compression technique and Huffman coding, to encode the bitstream. The complexity of the MVDR+PP decoder is $\mathcal{O}\left(N_\omega N_t K M^2\right)$.

3.4. Iterative sources reconstruction

This method is described in [14] and denoted ISS using Iterative Reconstruction (ISSIR) in the following. The spectrograms $\mathbf{S}_m$ are quantized uniformly on a logarithmic scale to yield $\bar{\mathbf{S}}_m$ and an *activity map* W_m is computed as:

$$W_m(\omega,t) = \begin{cases} 1 & \text{if } \frac{\mathbf{S}_m(\omega,t)}{\sum_{m'}\mathbf{S}_{m'}(\omega,t)} > T_{\text{act}} \\ 0 & \text{otherwise} \end{cases}.$$

where T_{act} is an *activity threshold*. The side information $\left\{\bar{\mathbf{S}}_m, W_m, A\right\}$ is compacted using the bzip2 coder. At the decoder, the method initially sets $\hat{S}_m = \sqrt{\bar{\mathbf{S}}_m}\frac{Y_0}{|Y_0|}$ and then iterates N_I times the following operations. First, the sources are reconstructed and the error between left and right-hand sides of (1) using $\hat{S}_m$ is computed. Second, this error is uniformly distributed to the sources based on W_m. Finally, a Griffin and Lim iteration [7] is performed. Two variants of the method are presented, one with single resolution (2048 point STFT) denoted "ISSIR-Single" and one with dual-resolution STFT (256 and 2048 points) with transient detection, denoted "ISSIR-Dual". The complexity of the ISSIR decoders are: $\mathcal{O}\left(N_\omega N_t K M N_I\right)$ for ISSIR-Single and $\mathcal{O}\left(\left(N_{\omega,1}N_{t,1} + N_{\omega,2}N_{t,2}\right)KMN_I\right)$ for ISSIR-Dual, where $N_{\omega,i}N_{t,i}$ stands for the number of TF bins for each transform.

4. EVALUATION

The main goal of this study is to proceed to a comparison of the objective performance of the four techniques for ISS presented in Section 3, as well as to identify some promising research directions. To this purpose, a set of 14 excerpts sampled at 44.1kHz or 48kHz from the Quaero database[2] was gathered along with all their constitutive audio objects. Each excerpt is approximately 30s long and composed of 5 to 10 objects. Two mixing scenarios were considered: linear instantaneous and convolutive mixtures using fixed short filters of order 200. In any case, the azimuth for all objects were different and separated by at least 5°. The filters used for mixing were the Head Related Transfer Functions (HRTF) described in [1].

The metrics considered are the Signal to Distortion Ratio (SDR, in dB) of BSSEval [16] and the Perceptual Similarity Measure (PSM, between 0 and 1) of PEMO-Q [8]. Whereas SDR is widely used for the objective evaluation of the performance of source separation algorithms, PEMO-Q is mostly used within the audio coding community. Even if both metrics are intended to be related to the perceptual accuracy of the estimates, SDR has been acknowledged as not sufficiently taking musical noise into account [9], which motivates the combined use of PSM. Further quantities of interest are the encoding and decoding times for all excerpts.

All techniques were then run at various levels of quality, in order to compare their respective rate-performance curves. Since objective performance appears to be highly dependent on the excerpt considered, all results are compared as in [9] to those of the oracle MMSE estimate (2), which makes it possible to compare metrics across different excerpts. Of course, performance of (2) should not be understood as a bound whatsoever, since it is limited by its use of the phase of the mixture to compute estimates, a limitation which is overcome by some of the techniques under study. Instead, it must be seen as a handy way to make the performance of all methods as independent as possible from the excerpt considered. Performance of the MMSE Oracle estimate for both instantaneous and convolutive mixtures can be found in Fig. 2.

For a given technique, a given excerpt and a given bitrate, the estimated audio objects were first compared to the original. Second, the obtained scores were averaged so as to obtain the metric for this technique, excerpt and quality. Third, the metrics obtained by the oracle estimate on the same excerpt were subtracted so as to obtain the differential metric $\Delta(\text{method}, \text{excerpt}, \text{bitrate})$. Finally, for a given method, the Δ of all excerpts were merged together and the scatter plot $(\text{bitrate}, \Delta)$ was smoothed to obtain one single rate-performance curve. With 10 quality settings for each of the 6 techniques considered on the 14 excerpts using both instantaneous and convolutive mixing, this evaluation involved computing metrics for 1680 groups of 5 to 10 stems. All results for both instantaneous and convolutive mixtures are given in Fig. 2.

From this figure, many noticeable facts may be underlined. First, the availability of a side information Θ yields a strong improvement of the performance obtained by source separation methods both in terms of SDR and PSM. Perceptually, original and estimated audio objects are hardly distinguishable, provided the bitrate is sufficient. This fact is

[1] http://www.bzip2.org

[2] www.quaero.org

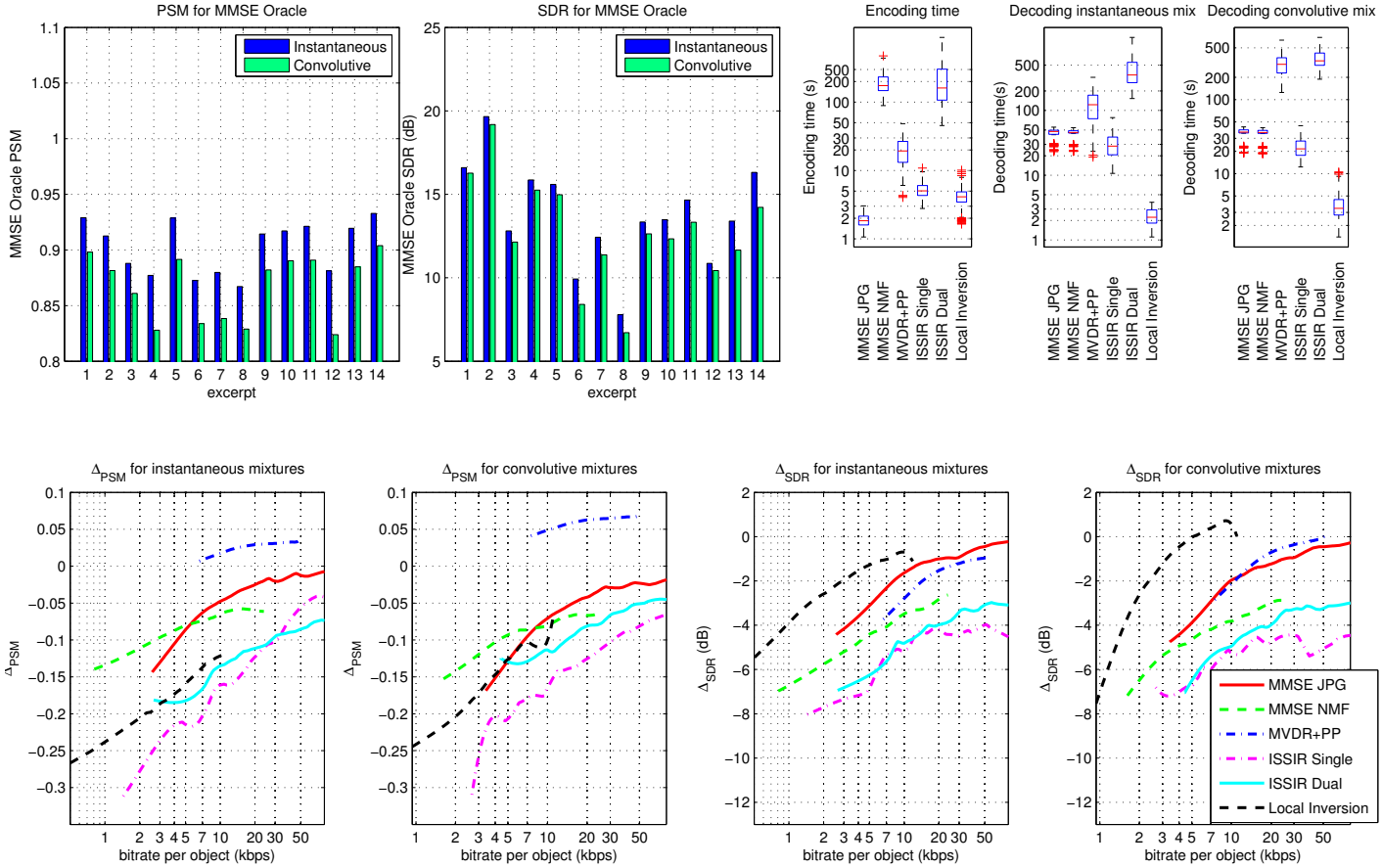

Fig. 2. Results for instantaneous and convolutive mixtures.

confirmed by objective results obtained by all methods above $5 - 10$kbps. ISS thus indeed yields estimates of sufficient quality for active listening applications. Such bitrates are quite low compared to those achieved by traditional perceptual coders.

Concerning the difference between the methods, it can be noticed that the performance of MMSE techniques is bounded by their limitative assumptions on the signals. Indeed, they use the phase of the mixture for estimation, as well as the assumption that energies of the objects are additive, which may not be true in many cases. These issues are addressed by other methods. Through its iterative reconstruction, ISSIR implements phase consistency of TF representations and MVDR+PP can be seen to outperform even Oracle MMSE in terms of PSM thanks to its exploitation of spatial constraints in the model. Local inversion can be seen to outperform all other methods in terms of SDR. Still, it yields some musical noise, due to the TF bins set to 0 in the reconstruction, a fact which is highlighted in the PSM results. Performance of ISSIR is seen to be a bit lower than other methods, but it is mainly because its current implementation does not benefit from the availability of several channels in the mixture. A noticeable fact about ISSIR and MMSE methods is that they are operational for mono mixtures as well as for mixtures setting the same mixing filters to different objects, which is not the case for Local Inversion nor for MVDR+PP methods. Concerning computing time, Local Inversion is by far the most computationally efficient method in both instantaneous and convolutive mixtures. While MMSE and ISSIR Single decoding methods can be seen to be possibly implemented in real time in any case, MVDR+PP currently requires more optimization for convolutive mixtures, as does ISSIR Dual for both instantaneous and convolutive mixtures.

Several interesting directions for research can be outlined. First, most techniques share a very similar side information. Indeed, a coded version of the spectrogram of the mixtures or a related quantity is often required at the decoder. Hence, we are currently exploring crossovers between methods on this point, particularly through the use of the NMF/JPG lossy compression approach. Second, constraints on spatial information as in MVDR+PP or on the phases as in ISSIR do yield a sensible improvement of the perceived quality. The corre-

sponding ideas may be transferred to other methods as well.

5. CONCLUSION

In this paper, we provided a unified presentation of Informed Source Separation. The challenge of ISS is to permit recovery of audio objects within their mixture through the use of a small side information Θ. Several methods for this purpose were presented that share the same common principle: Θ is chosen as the set of all parameters needed to perform source separation at the decoder. An extensive evaluation of the six techniques presented was done that allowed for the first time to compare the methods using the same corpus and the same metrics.

Noticeable results of this evaluation include the fact that ISS is highly efficient for recovering audio objects at a low bitrate ($5 - 15$kbps per object) and also that many crossovers between existing methods can be considered. It is interesting to mention that any audio separation method whose parameters can be encoded in a small amount of bits may be an interesting candidate for ISS.

6. REFERENCES

[1] V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano. The CIPIC HRTF Database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'2001)*, pages 99–102, New Paltz, New York, USA, October 2001.

[2] C. Avendano. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'2003)*, pages 55 – 58, October 2003.

[3] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408 – 1418, August 1969.

[4] N.Q.K. Duong, E. Vincent, and R. Gribonval. Underdetermined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1830 –1840, September 2010.

[5] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen. Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding. In *124th Audio Engineering Society Convention (AES 2008)*, Amsterdam, Netherlands, May 2008.

[6] S. Gorlow and S. Marchand. Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'2011)*, pages 309 –312, October 2011.

[7] D.W. Griffin and J.S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[8] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902 –1911, November 2006.

[9] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937 – 1949, 2012.

[10] H.O. Oh, Y.W. Jung, A. Favrot, and C. Faller. Enhancing Stereo Audio with Remix Capability. In *AES 129th Convention Preprint 8290*, San Francisco, CA, USA, November 2010.

[11] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Informed source separation: source coding meets source separation. In *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, New Paltz, New York, USA, October 2011.

[12] M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1721 – 1733, August 2011.

[13] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1464–1475, 2010.

[14] N. Sturmel and L. Daudet. Informed source separation using iterative reconstruction. arXiv:1202.2075v1.

[15] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet. Linear mixing models for active listening of music productions in realistic studio conditions. In *132th AES convention, Budapest, in press*, 2012.

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