INFORMED MULTIPLE-F0 ESTIMATION APPLIED TO MONAURAL AUDIO SOURCE SEPARATION

Dominique Fourer

LaBRI – CNRS
University of Bordeaux 1, France
dominique.fourer@labri.fr

Sylvain Marchand

Lab-STICC – CNRS
University of Brest, France
sylvain.marchand@univ-brest.fr

ABSTRACT

This paper proposes a new informed source separation technique which combines music transcription with source separation. The presented system is based on a coder/decoder configuration where a classic (not informed) multiple-F_0 estimation is applied on each separated source signal assumed known at the coder before the mixing process. Thus, the extra information required to recover the reference transcription of each isolated instrument is computed and inaudibly embedded into the mixture using a watermarking technique. At the decoder, where the original source signals are unknown, instruments are separated from the mixture using the informed transcription of each source signal. In this paper, we show that a classic (non-informed) F_0 estimator can be used to reduce the amount of bits necessary to transmit the exact transcription of each isolated instrument.

1. INTRODUCTION

Audio source separation aims to recover the original source signals which compose a mixture. The under-determined case, where the number of sources K is greater than the number of observations remains challenging and cannot efficiently be processed by independent component analysis (ICA). This particularly difficult configuration is often treated using signal sparse decomposition, thus source signals are separated using the prior knowledge about the source structure and its properties (e.g. orthogonality, harmonicity) [1, 2]. Unfortunately, the performance of the blind approach is often insufficient for the demanding applications and the informed approach has been considered in more recent works.

Informed source separation (ISS) [3] proposes to use directly the original source signals as extra information. This approach addresses the source separation problem in a coder/decoder configuration. At the coder, the extra information is estimated from the separated source signals before the mixing process and is inaudibly embedded into the signal mixture using watermarking. At the decoder, this information is decoded and used to assist the separation process. In spite of good results, ISS methods are comparable to audio coding and lossy audio compression where the resulting quality depends on the amount of transmitted extra information. In this particular case, the quality is related to the capacity of the watermarking technique and a trivial solution may consist in embedding directly the compressed original source signals into the mixture signal with the best distortion-rate ratio.

Symbolic information approaches addressed in [4, 5] use the aligned MIDI to assist the separation process. These techniques assume that each source signal follows a harmonic source model which is the most important part of the considered mixture. This assumption is often verified in tonal music for the pitched instruments [6] but cannot be applied to percussive instruments. Estimating a reliable and aligned transcription of each separated instrument is crucial for the separation quality and remains challenging. However, this point is often omitted by the proposed score-informed separation methods which assume the exact score is exactly known.

Here, we propose a fully automatized source separation system based on a coder/decoder scheme common to ISS methods. The aligned MIDI is estimated from the separated instruments source signals using a multiple-F_0 estimator. Thus the extra information required to recover each separated source transcription at the decoder is computed and inaudibly embedded into the mixture. At the decoder, the score is estimated from the mixture and corrected using the extra information. Finally, the sources are separated by a state-of-art score-informed source separation method.

This paper is organized as follows. Music transcription framework and notations are described in section 2. Section 3 proposes a technique for informed multiple-F_0 estimation. Experiments and results are presented in section 4. Finally, conclusions and future works are discussed in section 5.

2. MUSIC TRANSCRIPTION FRAMEWORK

We consider here a monaural discrete linear instantaneous sound mixture composed of K distinct source signals expressed as:

$$x[n] = \sum_{k=1}^{K} s_k[n] + r[n] \qquad (1)$$

where $r[n]$ is a residual noise signal resulting from source modeling and the watermarking process of the proposed system. Source separation aims to recover each source signal $s_k[n]$ from the mixture $x[n]$. In this study, each source signal $s_k[n]$ is assumed to be a music pitched instrument which follows a harmonic source model as described below.

2.1. Harmonic Source Model

It has been shown that pitched instrument sounds have a specific frequency structure [6]. Thus, each individual note is composed of one fundamental frequency, also called F_0, and several overtone partials. Quasi-harmonic sounds have overtone partials which are integer multiple of the fundamental frequency. However, for natural sounds, instruments may have shifted overtones due to inharmonicity and timbre. Thus, harmonic source signals can be modeled as a sum of complex exponentials according to Fourier theorem including inharmonicity and polyphony which can be expressed for a local frame analysis with the stationary assumption as follows:

$$s_k[n] = \sum_{l=1}^{L_k} \sum_{h=1}^{H_l} a_{k,h,l} \exp\left(j\left(2\pi\delta_\beta(h)hF_{k,l} \cdot n + \phi_{k,h,l}\right)\right) \tag{2}$$

where $j^2 = -1$ and $\delta_\beta(h) = \sqrt{1+h^2\beta}$ is the inharmonicity factor depending on β parameter according to Fletcher and Rossing [6]. L denotes the polyphony (the number of simultaneous notes per source) and H is the number of harmonics per note. Equation (2) uses the stationary sinusoidal model where a and ϕ denote respectively the instantaneous amplitude and initial phase. The F parameter corresponds to the perceived fundamental frequency which is related to the music pitch according to the following expression:

$$P(f) = \left[P_{\text{ref}} + 12\log_2\left(\frac{f}{f_{\text{ref}}}\right)\right] \tag{3}$$

where $[.]$ denotes the round-to-nearest-integer operation. $P(f_{\text{ref}}) = P_{\text{ref}}$ is the reference pitch of frequency f_{ref}. The MIDI specifications define $P(440) = 69$ corresponding to the note A_4. Most frequencies of pitched instruments are in the [27.5Hz, 7040Hz] range corresponding respectively to the notes A_0 and A_8.

2.2. Classic Multiple-F_0 Estimation

Music transcription aims to estimate the fundamental frequency of each harmonic source present in a mixture $x[n]$. This task can be efficiently processed in the monophonic case by time-domain techniques using autocorrelation or difference functions. The YIN algorithm [7] which is a reference state-of-art method is robust to noise, computationally efficient and obtains good results for the monophonic case ($L = 1$). The polyphonic case where $L > 1$ is still an open issue due to overlapped components. Furthermore, it is also difficult to estimate the polyphony and to separate the corresponding score of each individual instrument.

Efficient state-of-art methods for music transcription estimate the score and the polyphony simultaneously without separation [8, 9]. The best methods evaluated at the MIREX[1] 2011 reaches about 68% of accuracy for the multiple-F_0 estimation task on the evaluation database. The approach proposed by Yeh and Roebel [9] was implemented and integrated into the proposed system of this study. It was used for the transcription of each isolated source. This method obtained particularly good results in our experiments and was shown to be computationally less expensive than Monte Carlo Markov Chain (MCMC) though it is based on the generation of F_0 candidates. Thus, each candidate is evaluated by a score function depending on inharmonicity, spectral envelope and time synchronicity as detailed in [9].

3. INFORMED MULTIPLE-F_0 ESTIMATION

3.1. Overall Method Description

The proposed system uses a coder and a decoder which are respectively described in Fig. 1 and Fig. 2.

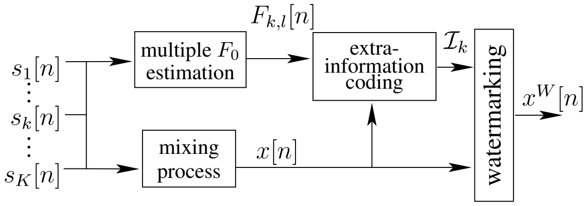

Fig. 1. The coder.

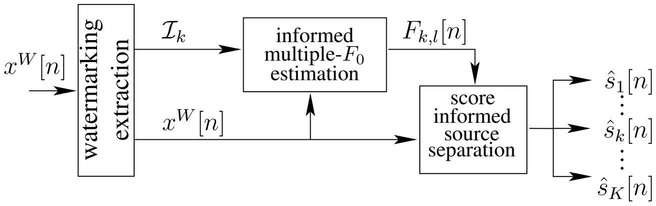

Fig. 2. The decoder.

At the coder, we assume each $s_k[n]$ signal is available before the mixing process. The score transcription of each isolated $s_k[n]$ is computed using a multiple-F_0 estimator algorithm based on [9]. Thus, the extra information required to recover each isolated score from the mixture is coded using the algorithm scheme of Fig. 3 and Fig. 4. The resulting binary code is inaudibly embedded into the mixture signal using the watermarking technique described in [10].

[1] Music Information Retrieval eXchange homepage: `http://www.music-ir.org/mirex/wiki/2011:MIREX2011_Results`

At the decoder, the signals $s_k[n]$ are unknown. The extra information extracted from the mixture is used to correct errors and to separate the music transcription resulting from the multiple-F_0 estimation algorithm applied on the watermarked mixture $x^W[n]$. Finally, a score-informed source separation method [5] is used to estimate the separated source signals $\hat{s}_k[n]$ from $x^W[n]$.

3.2. Informed Score Estimation

According to equation (2), each source signal $s_k[n]$ corresponds to a set of active notes which are related to a fundamental frequency. In our method, each note is coded as a MIDI pitch using equation (3). The pitch resolution allows to represent each note with 7 bits of information (using a binary code) and is sufficient enough for the score-informed separation algorithm applied at the decoder.

As discussed in section 2, a multiple-F_0 estimator can approximate the set of all active notes played in a mixture but cannot separate the corresponding score of each isolated instrument without prior information.

Thus, the proposed coder/decoder configuration has two goals. First, it aims to recover the reference set of all instantaneous active notes (for all the instruments) present in the mixture. Second, it aims to affect each note to each corresponding instrument as required by the separation process. The corresponding extra information required to assist a classic multiple-F_0 estimator is computed at the coder using the algorithm illustrated in Fig. 3 and Fig. 4.

Let us use the following notations for the proposed method description. $\Pi_k^{(t)}$ denotes the instantaneous set of all active notes for the source k and $|\Pi_k^{(t)}| \in [0, L_k]$ is the cardinality of this set. This corresponds to the number of simultaneously active notes for the source k located at instant t. $\Omega^{(t)} = \bigcup_{k=1}^{K} \Pi_k^{(t)}$ denotes the overall set of all active notes (where each element is unique) and $\hat{\Omega}^{(t)}$ is the set of estimated notes resulting from a classic multiple-F_0 estimation applied on the mixture $x[n]$. According to Fig. 1, the reference $\Pi_k^{(t)}$ is estimated at the coder by the multiple-F_0 estimator applied on each separated signal $s_k[n]$.

3.2.1. Coder

The proposed algorithm computes a binary code $\mathcal{I}^{(t)}$ from the generated *insertion/suppression* or *prediction* operations which must be applied to recover $\Omega^{(t)}$ from $\hat{\Omega}^{(t)}$. This corresponds to the common errors committed by existing music transcription systems and which are used as an evaluation metric in [11]. During the algorithm procedure, $\tilde{\Omega}^{(t)}$ is computed and corresponds to the set of active notes which can be computed using $\mathcal{I}^{(t)}$. Thus, the coding process is terminated when $\tilde{\Omega}^{(t)} = \Omega^{(t)}$. At a first step, $\Omega^{(t)}$ is compared to $\Omega^{(t-1)}$ and 1 bit is used to inform if the decoder has to consider the prediction: $\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t-1)}$. In this particular case, all the active notes are sustained and $\hat{\Omega}^{(t)}$ is ignored. Otherwise, each *insertion* and each *suppression* operation required to compute $\tilde{\Omega}^{(t)}$ from $\tilde{\Omega}^{(t-1)}$ is marked as *accepted* or *rejected* using 1 bit. Fig. 3 and Fig. 4 uses the following code convention: 1 is used for *accepted* and 0 otherwise. Thus each bit is concatenated to $\mathcal{I}^{(t)}$. When all transformation operations are treated to obtain $\tilde{\Omega}^{(t)}$, it is compared with $\Omega^{(t)}$. If $\tilde{\Omega}^{(t)} \neq \Omega^{(t)}$, all the missing *insertion* and *suppression* transformation operations are directly coded into $\mathcal{I}^{(t)}$ using 7 bits per F_0 candidate. One can deduct this coding step increases heavily the size $\mathcal{I}^{(t)}$ and can be avoided with a more accurate estimator.

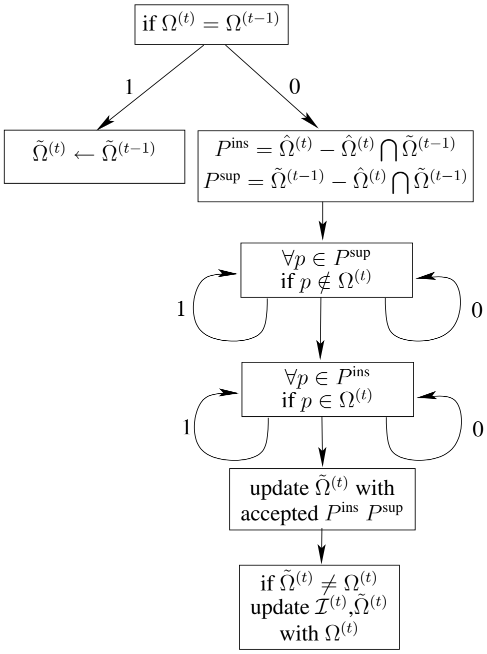
if $\Omega^{(t)} = \Omega^{(t-1)}$
1
0
$\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t-1)}$
$P^{\text{ins}} = \hat{\Omega}^{(t)} - \hat{\Omega}^{(t)} \bigcap \tilde{\Omega}^{(t-1)}$
$P^{\text{sup}} = \tilde{\Omega}^{(t-1)} - \hat{\Omega}^{(t)} \bigcap \tilde{\Omega}^{(t-1)}$
$\forall p \in P^{\text{sup}}$
if $p \notin \Omega^{(t)}$
1
0
$\forall p \in P^{\text{ins}}$
if $p \in \Omega^{(t)}$
1
0
update $\tilde{\Omega}^{(t)}$ with
accepted P^{ins} P^{sup}
if $\tilde{\Omega}^{(t)} \neq \Omega^{(t)}$
update $\mathcal{I}^{(t)}$,$\tilde{\Omega}^{(t)}$
with $\Omega^{(t)}$

Fig. 3. Algorithm scheme used to compute $\mathcal{I}^{(t)}$ which is required to obtain $\tilde{\Omega}^{(t)}$ from $\hat{\Omega}^{(t)}$. P^{ins} and P^{sup} denote respectively the set of note candidates which have to be inserted or suppressed.

The set of active notes per source $\tilde{\Pi}_k^{(t)}$ is coded with a similar strategy. One bit is necessary to inform the decoder to use the prediction for the considered source, $\tilde{\Pi}_k^{(t)} \leftarrow \tilde{\Pi}_k^{(t-1)}$. Otherwise, *insertion* and *suppression* operations are coded and applied to obtain $\tilde{\Pi}_k^{(t)}$. Finally, $\mathcal{I}^{(t)}$ is inaudibly embedded into the mixture signal using the desired watermarking technique.

3.2.2. Decoder

At the decoder, where the source signals $s_k[n]$ are unknown, $\mathcal{I}^{(t)}$ is exactly recovered from $x^W[n]$ by the watermark extraction. The algorithm starts with $\tilde{\Omega}^{(0)} = \tilde{\Pi}_k^{(0)} = \emptyset$. $\hat{\Omega}^{(t)}$ is computed from $x^W[n]$ using the same multiple-F_0 estimator. The *insertion* and *suppression* transformation operations

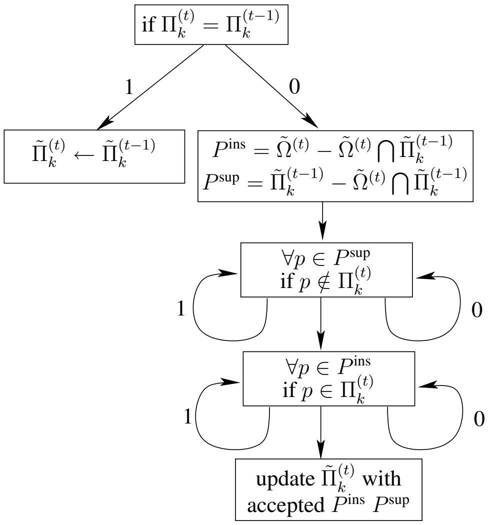
if $\Pi_k^{(t)} = \Pi_k^{(t-1)}$
1
0
$\tilde{\Pi}_k^{(t)} \leftarrow \tilde{\Pi}_k^{(t-1)}$
$P^{\text{ins}} = \tilde{\Omega}^{(t)} - \tilde{\Omega}^{(t)} \bigcap \tilde{\Pi}_k^{(t-1)}$
$P^{\text{sup}} = \tilde{\Pi}_k^{(t-1)} - \tilde{\Omega}^{(t)} \bigcap \tilde{\Pi}_k^{(t-1)}$
$\forall p \in P^{\text{sup}}$
if $p \notin \Pi_k^{(t)}$
1
0
$\forall p \in P^{\text{ins}}$
if $p \in \Pi_k^{(t)}$
1
0
update $\tilde{\Pi}_k^{(t)}$ with
accepted P^{ins} P^{sup}

Fig. 4. Algorithm scheme used to compute $\mathcal{I}^{(t)}$ which is required to obtain $\tilde{\Pi}_k^{(t)}$ from $\tilde{\Omega}^{(t)}$.

are computed and corrected using $\mathcal{I}^{(t)}$ and the inverse of the coding procedure is applied. The corrected transformation operations ensure to verify $\tilde{\Omega}^{(t)} = \Omega^{(t)}$ and $\tilde{\Pi}_k^{(t)} = \Pi_k^{(t)}$ at the end of the decoding process.

Due to signal differences between $x[n]$ and $x^W[n]$, $\hat{\Omega}^{(t)}$ may differ from the one computed at the coder, however the prediction mechanism integrated in this coding strategy increase the robustness of the method. This ensure the estimator is only used to introduce new note candidates at each onset or offset event. Thus all estimation errors do not have to be systematically coded and corrected during sustain periods.

3.3. Watermarking

The extra information $\mathcal{I}^{(t)}$ is inaudibly embedded using the technique presented in [10]. This method inspired from Quantization Index Modulation (QIM) [12] is based on the Modified Discrete Cosine Transform (MDCT) coefficients quantization. We chose this method for its high perceptual quality and capacity. Furthermore, this watermarking technique is frame-based and can theoretically allow real-time decoding for real-time applications. However this feature was not exploited during our experiments due to the low amount of embedded extra information. Unfortunately, this technique is not robust to lossy compression thus it limits the field of application to lossless formats (e.g. FLAC, WAVE).

3.4. Score-Informed Source Separation

After the decoding process where the score of each separated signal is estimated and quantized on the MIDI pitch scale, source signals are separated using [5]. This method based on Non-negative Matrix Factorization (NMF) approximates the magnitude spectrogram of the mixture denoted X of size $F \times T$ by a product between two positive matrices as:

$$X \approx \hat{X} = WH \tag{4}$$

where $W \in \mathbb{R}^{F \times R}$ and $H \in \mathbb{R}^{R \times T}$ ($FR + RT \ll FT$) correspond respectively to the constrained dictionary and time activations. This method uses a dictionary composed of harmonic time-dependent parametric atoms and uses $\Pi_k(t)$ as a prior to constrain the matrix H of time activation for each source signal. This method was shown to obtain good results comparable to Probabilistic Latent Component Analysis (PLCA) based algorithms [4].

4. EVALUATION

For these experiments, we use two musical pieces of 6 seconds sampled at 44.1kHz. The first piece is composed of 3 instruments: flute, piano, and contrabass. The second one is composed of 4 instruments: Hammond organ, piano, contrabass, and drum. All instruments except the drum were transcribed at the coder using our implementation of [9]. During experiments, we used time frames of 185ms (FFT length of 8192 samples) with an 75% overlap. The transcription returned by this estimator were post-processed with a reconstruction algorithm which suppresses the F_0 candidates with a duration lower than 4 frames and reconnects the F_0 trajectory. The same algorithm with the exact identical parameters is applied on the mixture during the decoder process. For the separation applied at the decoder, the magnitude spectrogram was computed with time frames of 46ms (FFT length of 2048 samples) and results were obtained after 5 iterations (see [5] for the update rules details of the NMF).

4.1. Informed Multiple-F_0 Estimation

Table 1 shows the overall amount of bits used to assist the score estimation at the decoder. The classic multiple-F_0 estimator combined with our coding algorithm achieves to reduce the amount of the extra information required to obtain the exact reference transcription. The resulting code computed without estimator ($\hat{\Omega}^{(t)} = \emptyset$) requires a lower amount of bits than the reference MIDI file.

	Piece 1	Piece 2
Proposed (with estimator)	2415 bits	2353 bits
Proposed (no estimator)	2582 bits	2542 bits
MIDI file	5424 bits	6120 bits

Table 1. Amount of transmitted information used at the decoder for informed transcription and applied to score-informed source separation.

4.2. Quality Evaluation of Source Separation

Table 2 and 3 show objective quality measures computed between the original and the estimated sources signals using BSS_Eval toolbox [13]. Thus, the performance is assessed with signal to distortion ratio (SDR), signal to interference ratio (SIR) and signal to artifact ratio (SAR) all defined in [13]. The informal listening tests show an acceptable listening[2] quality with perceptible artifacts. Our experiments show a negligible effect of the watermarking process on separation performance due to a low bitrate requirement (about 0.4kbps).

	SDR(dB)	SIR(dB)	SAR(dB)
flute	14.97	24.10	15.55
piano	6.55	22.63	6.68
bass	10.54	33.30	10.57

Table 2. Quality results for the piece 1.

	SDR(dB)	SIR(dB)	SAR(dB)
B3	4.89	11.09	6.41
piano	3.56	14.29	4.10
bass	2.58	5.72	6.49

Table 3. Quality results for the piece 2.

5. CONCLUSION AND FUTURE WORK

We have presented a score-informed source separation system for harmonic instruments which operates at a very low bitrate. We show that the informed approach can be combined with a classic multiple-F_0 estimator to reduce the amount of extra information. In our experiments, a gain of approximately 7% was observed using the proposed estimator for the informed transcription. Thus, this work can be considered as a proof of concept which has to be further investigated. In fact, the F_0 estimator applied on the mixture is independent from the proposed coding algorithm thus it is probably not optimized for a minimal bitrate. A future work will consist in evaluating the relationship between the estimation errors and the extra information requirement for correction and separation. The proposed framework has the advantage to be flexible enough to be combined with newer F_0 estimators, source separation techniques, and watermarking methods.

6. ACKNOWLEDGMENTS

Thanks go to R. Hennequin and J. Pinel for their implementations of the methods detailed respectively in [5] and [10]. This research was partly supported by the French ANR DReaM project (ANR-09-CORD-006).

7. REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in *IEEE TSP*, 2004, vol. 52, pp. 1830–1847.

[2] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," in *IEEE WASPAA*, Oct. 2001, pp. 83–86.

[3] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for single-channel audio source separation," in *IEEE ICASSP*, Apr. 2009, pp. 101–104.

[4] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S Abel, "Evaluation of a score-informed source separation system," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Aug. 2010, pp. 219–224.

[5] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *IEEE ICASSP*, May 2011, pp. 45–48.

[6] N. F. Fletcher and T. D. Rossing., *The Physics of Musical Instruments*, Springer-Verlag, 1998.

[7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] A. Klapuri, "Multiple fundamental frequency estimation by harmonicity and spectral smoothness," *IEEE TASLP*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[9] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE TASLP*, vol. 18, no. 6, pp. 1116–1126, 2010.

[10] J. Pinel, L. Girin, C. Baras, and M. Parvaix, "A high-capacity watermarking technique for audio signals based on MDCT-domain quantization," in *Int. Congress on Acoustics*, Oct. 2010.

[11] Graham E. Poliner and Daniel P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP JASP*, Oct. 2006.

[12] B. Chen and G.W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE TIT*, vol. 47, no. 4, pp. 1423–1443, May 2001.

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

[2] Sounds results are available on-line at: `http://www.labri.fr/perso/fourer/publi/EUSIPCO12/`