INFORMED SOURCE SEPARATION OF UNDERDETERMINED INSTANTANEOUS STEREO MIXTURES USING SOURCE INDEX EMBEDDING

Mathieu Parvaix & Laurent Girin

Grenoble Laboratory of Images, Speech, Signal and Automation (GIPSA-lab)
CNRS UMR 5216 , Grenoble Institute of Technology, Grenoble, France
Mathieu.Parvaix,Laurent.Girin@gipsa-lab.inpg.fr

ABSTRACT

In this paper, we address the issue of underdetermined source separation of non-stationary audio sources from a stereo (*i.e.* 2-channel) linear instantaneous mixture. This problem is addressed with a specific coder-decoder configuration. At the coder, source signals are assumed to be available before the mixing is processed. A time-frequency (TF) analysis of each source enables to select the one or two predominant sources (among I>2) in each TF region, and a corresponding source(s) index code is imperceptibly embedded into the mix signals using a watermarking technique. At the decoder level, where the original sources signals are unknown, the extraction of the watermark enables to locally reduce the underdetermined configuration to an (over)determined configuration. Sources signals can then be estimated using a classical (over)determined separation technique. Thereby several instruments or voice signals can be separated from stereo mixtures, enabling separate manipulation of the source signals during restitution (*i.e.* remastering).

Index Terms— underdetermined source separation, watermarking, audio processing, speech processing, remastering.

1. INTRODUCTION

Source separation aims at recovering I unobserved source signals $s_i[n], i \in [1, I]$, from J observations of their mixtures $x_j[n], j \in [1, J]$. The *underdetermined* case, where $J < I$, is a particularly difficult configuration. It cannot be processed by Blind Source Separation (BSS) / Independent Components Analysis (ICA) methods developped for (over)determined mixtures ($J \geq I$) [1] [2]. However, it is of particular interest in audio processing since, in the typical mono or stereo configuration, many instruments and voices are to be separated from only one or two channels. This would enable to separately manipulate the different elements of the audio scene, e.g., modifying the volume, the color or the spatialization of an instrument, a process referred to as *active listening* or *remastering*.

To achieve underdetermined source separation, many relevant techniques are based on sparse time-frequency (TF) representations of signals [3] [4]. In [5] [6] we introduced the concept of Informed Source Separation (ISS), with a specific coder-decoder configuration corresponding to the distinct steps of signal production (e.g. music recording/mixing in studio) and signal restitution (e.g. audio-CD at home). In addition to the mixture signals at the separation level (so-called here the decoder), source signals are thus assumed to be available at the mixing level (so-called here the coder). Parameters are extracted from the source signals at the coder, and this extra information is imperceptibly embedded into the mixture signals using watermarking techniques. Extracting the watermark at the decoder enables an end-user who has no direct access to the original sources (but only to the watermarked mixture signals), to separate these sources from the mixture signals. As for BSS, different approaches exist for ISS, depending on the assumptions made on the source signals (mutual independence, sparsity) and on the mixture (instantaneous, anechoic, convolutive, over/under-determined). As a result, the side-information embedded into the mixture, and the way it is used for the separation process may differ for the different configurations. In [5] [6], a single-channel instantaneous mixture of (speech/music) source signals was processed. The embedded information consisted of TF prototypes of the sources, issued from matricial codebooks. In the present study, we focus on underdetermined source separation of stereo (2-channel) instantaneous mixtures of music signals. Since we aim at exploiting the spatial information, the side-information is here reduced to the indexes of the one or two predominant sources in each TF region, as provided by an analysis of the source signals at the coder. At the decoder, extracting the watermarked indexes enables to locally reduce the underdetermined system to an "artificial" (over)determined one. The separation is then processed by classical matrix inversion techniques.

This paper is organized as follows. The proposed method is described in Section 2. Results obtained for music mixtures are given in Section 3. Finally, some perspectives are presented in Section 4.

2. THE METHOD

Fig. 1 presents the diagram of the proposed stereo Informed Source Separation (Stereo-ISS) technique. Some of the functional blocks of this diagram are identical to those described in [6] and thus will not be detailed. The present paper rather focuses on the new techniques of source analysis and separation (blocks 4 and 11 of Fig.1). In this study, the mixing process (block 1) is a multiplication of the I-source vector with a time-invariant $2 \times I$ matrix.

2.1. MDCT decomposition and molecular grouping

The source signals of interest are voice/instrument signals playing a same piece of music (but recorded separately). They are non-stationary, with possibly large temporal and spectral variability, and they generally strongly overlap in the time domain. Using a time-frequency (TF) representation of audio signals has been shown to exhibit natural sparsity, i.e. much lower overlapping of signals in the TF domain, thus leading to sparsity-based separation methods [3] [4] [6]. As in [6], the Modified Discrete Cosine Transform (MDCT) is used as the TF decomposition, since it presents good energy concentration properties. This transformation is used in blocks 2, 2' and 8 of Fig. 1 while the corresponding inverse transform (IMDCT)

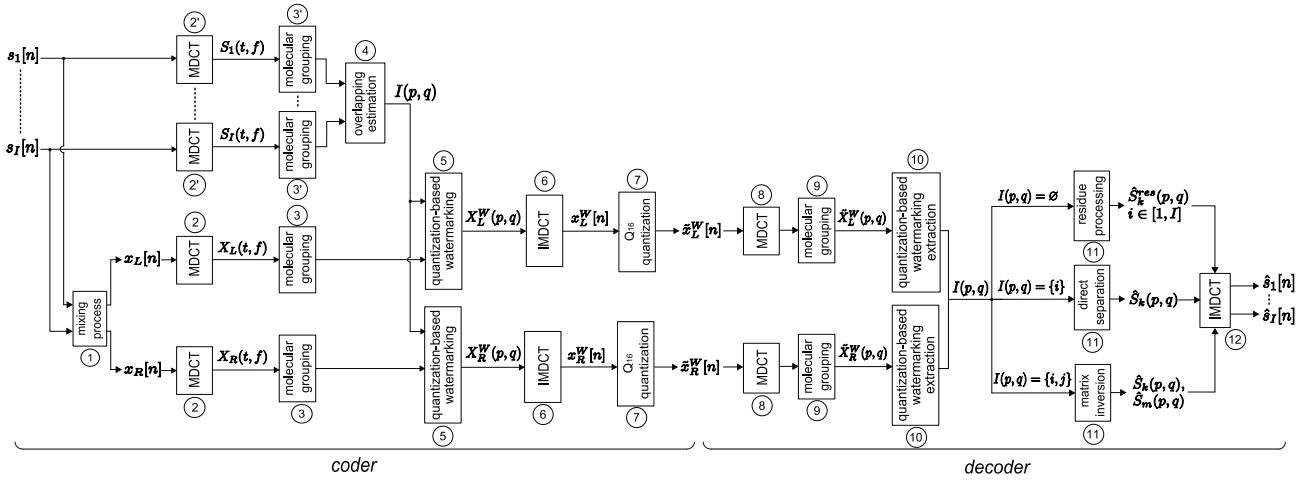

Fig. 1: Detailed structure of the proposed system for Stereo-ISS.

is used in blocks 6 and 12 to regenerate time-domain signals from MDCT coefficients. Since the MDCT is a linear transform, the source separation problem remains linear/instantaneous in the transformed domain. The MDCT is applied on time frames of W=2048 samples (46.5ms for a sampling frequency $f_s = 44.1$kHz), with a 50%-overlap between consecutive frames. This results in a matrix of MDCT coefficients $\mathcal{M}_x = \{X(f,t)\}$ of dimension 1024 frequency bins (denoted by f) by $L/1024$ time bins (denoted by t), where L is the overall length of the processed signal x. The frame length W is chosen to follow the dynamics of audio signals while providing a frequency resolution suitable for the separation.

As in [6], the side-information about source signals is embedded into the mixture signals using a watermarking technique applied on the MDCT coefficients (see Section 2.2). Because the embedding capacity of a single coefficient is too poor to embed all the side-information, neighboring coefficients are gathered into what is referred to as *molecules*. A molecule M^x_{pq} is the $F \times T$ sub-matrix of coefficients located at the so-called *molecular* frequency and time bins (p,q) in the TF plane (see Fig. 2). In the present study, the dimension of a molecule is 1×4. The forthcoming analysis, watermarking, and separation processes are all carried out at the MDCT molecule level.

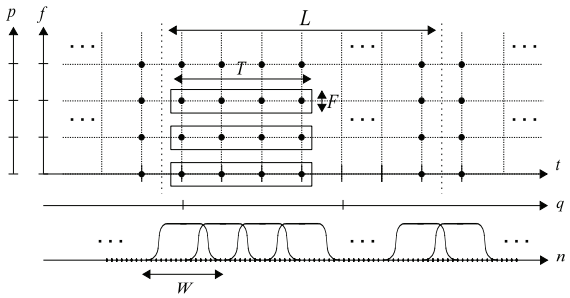

Fig. 2: Schematic representation of the time-frequency decomposition and molecular grouping.

2.2. Watermarking process

The watermarking technique is the same as the one presented in [5] [6]. It is inspired from the Quantization Index Modulation (QIM) of [7], with adaptation to the MDCT coefficients. Briefly speaking, the inserted message is carried by a quantization of the MDCT coefficients with a specific quantizer whose "sub-levels" are associated with watermark values. The capacity results from joint maximisation of a reference quantization step under inaudibility constraint, and minimisation of sub-level quantization step under robustness-to-noise constraint [5] [6]. Since we target the audio-CD application for our system, the noise under consideration results from the 16-bits linear quantization of the watermarked mixture samples (block 7 of Fig. 1), and the system is tuned so that the quantization of the MDCT coefficients at the coder (block 5) and at the decoder (block 10) provide the same result. This technique has been shown to provide a large capacity, up to about 150kb/s for music signals [6]. This is largely sufficient to embed the separation information that is used in the present study (see Sections 2.3 and 4).

2.3. Analysis of source contribution: how many and which?

As opposed to [5] [6], where "rich" descriptors of the source signals were embedded (the above-mentionned source prototypes), in the present method the watermark embedded into the mix signals contains very "primary" (but fundamental) information on the local composition of the mixture, as provided by the analysis stage of block 4: the number of predominant sources (*i.e.* sources with very high energy compared to the other sources) and their index within the sources set, for each molecular bin. Moreover, based on the sparsity of audio signals, we consider only three possible configurations for each molecular bin: 0) no source is present, 1) only one source is present, as in, e.g., [3], or 2) two sources are present, as in, e.g., [4]. Since we have a two-channel mixture, Cases 1) and 2) are (over)determined configurations and can then be processed by the process described in Section 2.4.

Case 0 results from music signals sparsity: many of the MDCT coefficients are close to zero. There is no need to process the separation in these TF regions of poor audio relevance[1]. For this reason, a

[1] Accordingly, the embedding capacity in these TF areas is too poor to

first step of the proposed method consists in thresholding the mixture signals: only the molecules with a sufficient energy are considered relevant for further processing, the remaining molecules being processed separately as residue.

To process Cases 1 and 2, the following power ratio is defined and calculated for each molecule $M_{pq}^{s_i}$ of each source s_i:

$$R_i(p,q) = \frac{\sum_{(f,t)\in\{P\times Q\}} |S_i(f,t)|^2}{\sum_{j\neq i} \sum_{(f,t)\in\{P\times Q\}} |S_j(f,t)|^2} \quad (1)$$

where $P \times Q = [(p-1)F, pF-1] \times [(q-1)T, qT-1]$. For each molecular bin, the two sources with the highest ratios, say s_k and s_m, are selected. Finally, Case 2 is reduced to Case 1 if $R_m(p,q) < \varepsilon R_k(p,q)$ with ε a small scalar factor (typically 0.05). In such case, only s_k is selected. The result of this analysis is encoded into the watermark $I(p,q)$ using very few bits (compared to the side-information processed in [6]), typically less than 4 bits for coding the 11 possible combinations with 4 sources (hence a bit-rate of approx. 40kb/s with 1×4 molecules). Note that, if the mixture is actually locally underdetermined, *i.e.* if more than 2 sources are actually present in a molecule, then only the two most energetic sources are taken into account in the separation process, the remaining sources being considered as noise.

2.4. Separation process

The separation process is carried out on the MDCT molecules of the mix signals (block 11 of Fig. 1), after the watermark $I(p,q)$ has been decoded (block 10). The $2 \times I$ mixing matrix $\mathbf{A} = \{a_{ji}\}$ is identical for every molecule, and is assumed to be known at the decoder, since it can easily be transmitted via watermarking, given the capacity range.

For each molecule, we have the three following possible processes depending on the $I(p,q)$ code, corresponding to the three cases of Section 2.3. If no source is present (Case 0), the molecule is considered as residue (that can be shared between sources, or simply left appart). If one source is present (Case 1), say s_k, the stereo mixture reduces to[2]:

$$\begin{bmatrix} X_L(p,q) \\ X_R(p,q) \end{bmatrix} = \begin{bmatrix} a_{1k} \\ a_{2k} \end{bmatrix} \begin{bmatrix} S_k(p,q) \end{bmatrix} \quad (2)$$

Hence, an estimate of $S_k(p,q)$ can be easily obtained by:

$$\hat{S}_k(p,q) = \frac{1}{a_{1k}} \tilde{X}_L^W(p,q) \text{ or } \hat{S}_k(p,q) = \frac{1}{a_{2k}} \tilde{X}_R^W(p,q) \quad (3)$$

or a combination of both. If two sources s_k and s_m are present (Case 2), the stereo mixture reduces to

$$\begin{bmatrix} X_L(p,q) \\ X_R(p,q) \end{bmatrix} = \begin{bmatrix} a_{1k} & a_{1m} \\ a_{2k} & a_{2m} \end{bmatrix} \begin{bmatrix} S_k(p,q) \\ S_m(p,q) \end{bmatrix} \quad (4)$$

and the corresponding estimate molecules are obtained by:

$$\begin{bmatrix} \hat{S}_k(p,q) \\ \hat{S}_m(p,q) \end{bmatrix} = \begin{bmatrix} a_{1k} & a_{1m} \\ a_{2k} & a_{2m} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{X}_L^W(p,q) \\ \tilde{X}_R^W(p,q) \end{bmatrix} \quad (5)$$

Finally, the source signals are reconstructed from the corresponding estimated source molecules by applying inverse MDCT (block 12 of Fig. 1).

3. EXPERIMENTS AND RESULTS

Tests have been processed with 44.1kHz-sampled music signals. We tested stereo mixtures of four sources: a female singer, a bass guitar, a piano, and drums (one track for the overall drum set). 10s-excerpts of two pieces of music played together by all four sources (recorded separately in studio conditions) were used. The quality of separated sources has been assessed by informal listening tests and performance measures, as defined in [8]. Basically, the source-to-distortion ratio (SDR) provides an overall separation performance criterion, the source-to-interferences ratio (SIR) measures the level of interferences from the other sources in a source estimate, and the source-to-artifacts ratio (SAR) measures the level of artifacts in a source estimate. We also provide the input SIR so that the difference between output SIR and input SIR measures the rejection power of the method.

3.1. Measurement of source overlapping

The overlapping or non-overlapping of source signals in the TF domain remains a critical issue for sparsity-based source separation techniques. In order to assess the relevance of the proposed separation approach on the presently used music signals, a measurement of source signals overlapping has been done. For each frequency bin, we first calculated the percentage of time frames with i significant sources, $i \in [0, I]$, after a selection of the 95% most energetic MDCT coefficients, as in [9]. A maximum of only 2 source signals out of 4 was shown to compose the mixture for more than 80% of the frames. Furthermore, if we look at the energy distribution of each source with respect to the rank of its power ratio (1), Table 1 shows that 97.1% (for drums) to 99.4% (for voice) of the energy of a source signal corresponds to the case where this source is within the two most energetic sources. All this justifies that at most two sources are considered in the separation process at the molecular level. Again, if three or four sources overlap, the third and fourth sources can be considered as noise. If the inverse matrix in (5) is not ill-dimensionned, the separation obtained by (5) provides good results even in this "noisy" case.

Table 1: Percentage of the overall energy of a source signal as a function of its local (molecular) energy rank in the mixture.

Rank	Bass	Singer	Drums	Piano
1	**82.3**	**95.4**	**78.5**	**94.7**
2	**16.0**	**4.0**	**18.6**	**4.3**
3	1.6	0.5	2.7	0.8
4	7.10^{-4}	3.10^{-4}	0.2	0.15

3.2. Separation results

Table 2 provides average results obtained with the proposed stereo-ISS method, for the separation of twenty-four 2×4 mixtures corresponding to all the possible permutations of the (normalized) basis vectors (*i.e.* columns) of the following matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 0.93 & 0.80 & 0.60 & 0.37 \\ 0.37 & 0.60 & 0.80 & 0.93 \end{bmatrix} \quad (6)$$

embed any extra information [6].

[2] In the following equations, $X(p,q)$ denotes a $1 \times T$ molecule of MDCT coefficients, because the configuration selection that drives the separation is defined at the molecule level. The equations are the same when applied separately to each MDCT coefficient of such a molecule. L/R denotes the left/right channel.

Moreover, a comparison was made with the underdetermined blind source separation process of [4] (further referred to as BZ-UBSS)[3]. In [4], for each TF point, two source signals (out of 4 here) are estimated by finding the linear combination of the two basis vectors that provides the shortest path from the origin to the observed data **x**. For example, in Fig. 3, the mixture vector **x** is a linear combination of sources 1 and 2. It can be noticed that such a geometrical method does not provide all the possible source combinations. For instance, if **x** is a linear combination of sources 1 and 3, this method will always return the spurious couples of sources (1,2) or (2,3). The watermark embedded in the proposed Stereo-ISS method fixes this issue. Note that the 24 permutations of **A** limit some potential artefacts due to the unfortunate case where the "impossible combinations" of BZ-UBSS would systematically concern the highest energy sources.

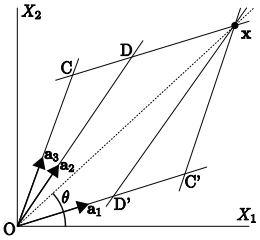

Fig. 3: Geometrical method of the *shortest path* from the origin to the data point **x** introduced in [4].

Table 2 shows that the Stereo-ISS method provides a very good separation of all four signals: SIR is increased from (-8.5)–(-0.5)dB at the input to 33.3–37.4dB at the output, hence a 36.2–44dB SIR improvement. This is confirmed by listening tests, that show a very good rejection of competing sources: each instrument is clearly isolated. Of course, the quality is not perfect (SDR/SAR range within 9.5–14.7dB), and some level of musical noise remains. However, the isolated source signals can be clearly enhanced or subtracted from the mixture signals, and in this case the musical noise appears to be largely masked by the mixture. Table 2 also reveals a large advantage of Stereo-ISS over BZ-UBSS, for every performance measure, demonstrating the benefit of using side-information. Accordingly, the audio quality of the separated signals is clearly better for the Stereo-ISS method. Sound samples can be downloaded at `http://www.icp.inpg.fr/~girin/Stereo-ISS-demo.rar`.

Table 2: Separation performances averaged over 24 mixtures (8 minutes).

Signals	input	Stereo ISS			BZ-UBSS		
	SIR	SDR	SIR	SAR	SDR	SIR	SAR
bass	-5.4	11.5	33.3	11.5	6.6	14.3	7.6
singer	-0.5	14.7	35.7	14.7	9.9	17.7	10.8
drums	-8.5	9.5	34.2	9.5	3.8	10.9	5.3
piano	-6.6	12.7	37.4	12.7	6.4	12.8	7.8

[3] with the mixing matrix **A** being assumed to be known in both cases for fair comparison (in [4], **A** is claimed to be accurately estimated by a clustering technique).

4. CONCLUSION

The ISS method described in this paper does not belong to classical source separation methods. Contrary to the BSS framework, source signals are available before the mix is processed, and specific applications such as active-listening from audio-CD are targeted. After the promising preliminary results of [6] in the single-channel case, the present paper shows significant advances in the 2-channel configuration. The simplicity of the side-information (and quite lower bitrate) compared to the source coding approach of [6] is compensated by an efficient exploitation of signals sparsity and spatial information. The side-information (including the single-shot transmission of the mixing matrix) allows to relax the "only one predominant source" assumption of [3] to 2 predominant sources, assumes that those 2 predominant sources are correctly selected (what cannot be done in [4]), and enables a very simple separation process. The combination of such simple approach with the previous source coding approach of [6] appears as a logical future extension of this work: such an hybrid system would be able to separate a quite large number of sources. For example, if 4 sources are locally predominant (hence overlapping) in a mixture of, say, 8 sources, 2 of them could be extracted by the coding approach of [6], and the remaining 2 others could be extracted with the present method, after subtraction of the first 2 decoded sources. Also, future work will deal with more complex types of mixture such as binaural mixtures.

5. REFERENCES

[1] J.F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley & Sons, 2001.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[5] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for single-channel audio source separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2009, pp. 101–104.

[6] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Trans. Audio, Speech, and Language Process.*, 2009, accepted.

[7] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[8] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, 2005.

[9] S. Araki, H. Sawada, and S. Makino, *K-means Based Underdetermined Blind Speech Separation*, pp. 243–270, in S. Makino and al. (Eds), Blind Source Separation, Springer, 2007.