A WATERMARKING-BASED METHOD FOR SINGLE-CHANNEL AUDIO SOURCE SEPARATION

Mathieu Parvaix[1]*, Laurent Girin*[1]*, Jean-Marc Brossier*[2]

Grenoble Lab. of Images, Speech, Signal and Automation (GIPSA-lab, DPC[1]/DIS[2])
CNRS UMR 5216 , Grenoble Institute of Technology, Grenoble, France
Mathieu.Parvaix,Laurent.Girin,Jean-Marc.Brossier@gipsa-lab.inpg.fr

ABSTRACT

In this paper, we address the issue of audio source separation with a single channel, i.e. the estimation of source signals from a single mixture of these signals. This problem is addressed with a specific configuration: source signals are assumed to be available before the mix is processed. We propose an original method that uses a watermarking technique to embed information about the source signals into the mix signal. Extracting this watermark enables an end-user who has no access to the original sources to separate these signals from their mixture. Thereby several instruments or voice signals can be segregated from a single piece of music to enable post-mixing processing such as volume control.

Index Terms— source separation, watermarking, audio system, speech processing.

1. INTRODUCTION

Source separation became in the past twenty years one of the most challenging problems in signal processing. It aims at estimating N original source signals from P observations of their mixture. In blind source separation (BSS) [1] [2], very few knowledge about the sources and the mixture process is available. It is clearly a complex issue, especially in the so-called *under-determined* case, where fewer observations than sources are available. In this study, we focus on audio source separation in a very specific configuration: we assume that a single observation of the (linear) mixing is available at the separation level (so-called here the decoder), but we also assume that source signals are available at the mixing level (so-called here the encoder). This is quite an original, and at first sight surprising, configuration in the source separation framework, but not unnatural, since in some applications mixing and demixing can be processed separately by cooperative users. For instance, we address here the audio-CD configuration: in a recording studio, the different tracks (corresponding to the different instruments and singing voice(s)) are recorded separately, and are then mixed in a very controlled way. But at home, an end-user has only the mix on the CD (actually, 2 stereo channels are available but they are generally very redundant and the exploitation of stereo is not considered here). The objective is there to enable the separation of the different elements of the audio scene, so that they can be manipulated separately (for example, the volume or the color of an instrument can be modified). This is an old dream of music lovers, that can be referred to as active listening.

In this paper, we propose a first approach to address this original problem at both coder and decoder levels. Figure 1 describes the principle of the proposed method (in the case of a mix of 2 signals). This method is based on the original combination of *source separation* with another major domain of signal processing, namely *watermarking*. Audio watermarking consists in embedding extra information within a signal in an inaudible manner. It is mainly dedicated to Digital Right Management (DRM). In the present study, we take advantage of the knowledge of source signals at the encoder to extract a set of descriptors from these signals to be embedded onto the mixing using a watermarking process. Defaults of the human hearing system are used to insert this information in MDCT coefficients of the mix signal using a quantization-based watermarking technique. At the decoder, the descriptors are extracted from the mixing, and then used for the separation process. Since the method exploits the knowledge of unmixed signals, the corresponding framework can be labelled as "informed source separation" (ISS), in opposition to BSS (note that, since only a single mix channel is used here, the method proposed in this paper is closer to Computational Audio Scene Analysis (CASA) [3] than to BSS, but the basic principle of ISS can easily be generalized to a multichannel framework).

This paper is organized as follows. In Section 2 a description of the proposed method and its application to audio signals is given. Results of speech signals separation are given in Section 3. Finally, some perspectives are presented in Section 4.

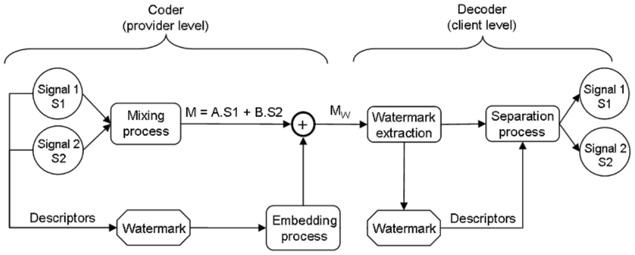


Fig. 1: Basic principle of the proposed post-mixing Watermarking Based Source Separation method.

2. THE METHOD

2.1. MDCT decomposition and molecular grouping

Since the different source signals that compose the additive mixture strongly overlap in the time domain, the first step of the proposed method is to find an appropriate representation in which the different sources are as segregated as possible. Also, sources can be non-stationary and have a large spectral variability, so that their respective contributions in the mix strongly depend on both time and frequency parameters. For all these reasons, a time-frequency (TF) decomposition of all signals is first applied at the encoder: the TF

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

representation of a given source signal is used to generate the corresponding source descriptors, that will be embedded into the TF representation of the mix signal. The Modified Discrete Cosine Transform (MDCT) [4] has been chosen as a TF decomposition, for its ability to concentrate the energy of audio signals into restricted regions of the TF plane. This property, already exploited in audio coding, is expected to be useful for source separation.

As further detailed in Section 2.3, the descriptors of source signals used for separation generally require a large embedding capacity from the host signal. Embedding such information into a single MDCT coefficient is not feasible. Thus, we propose to gather MDCT coefficients into what is referred here to as *molecules*, after [5] [6]. A *molecule* of neighboring MDCT coefficients is chosen to be the elementary support of the watermark information. It is also the elementary pattern from which the source signals descriptors used in the separation process are derived. In other words, the description, watermarking, and separation processes are all carried out at the molecule level. If m^s_{ij} denotes the entries of the MDCT decomposition matrix of a signal s, a molecule M^s_{pq} located at the coordinates (p, q) in the TF plane is defined by:

$$M^s_{pq} = \left\{m^s_{ij}\right\}_{\substack{i\in P=[(p-1)F+1,pF] \\ j\in Q=[(q-1)T+1,qT]}} \tag{1}$$

where T and F are the size of M^s_{pq} in time and frequency bins. The larger a mix signal molecule is, the larger is its watermarking capacity, and thus, the more information on source descriptors can be inserted into this molecule. Conversely, too large molecules cannot provide a good resolution for the separation of overlapped source signals. Therefore a trade-off between the watermarking capacity and the quality of source separation needs to be found when setting the values of T and F.

2.2. Watermarking process

Even though watermarking is used in an original way in ISS, it has to verify the following basic principles: being imperceptible to human ear, being exactly retrievable at the decoder, and being resistant to attacks. In the audio-CD application, we consider that the system does not have to cope with intentional attacks. The conversion to the audio-CD format (16-bits linear quantization of signal samples) of the watermarked mix is the only "attack" to be considered since it leads to a modification of the MDCT coefficients, and thus possibly of the watermark[1]. This explains the possibility to consider a high embedding rate, necessary to encode the source signal descriptors. For this reason, we use a quantization-based watermarking method inspired by [7], in which the inserted message is carried by a modification of quantization levels of the host signal. In the present study, this principle is applied to the MDCT coefficients of the mix signal. We take advantage of the fact that the accuracy of the MDCT coefficients of the mix signal is too "thin" for human ear. MDCT coefficients can be quantized with a quite "coarse" resolution without noticeable consequences on the quality of the resulting time signal. This principle is exploited in compression algorithms such as MPEG [8].

In the proposed method, the watermark is embedded into MDCT coefficients using two uniform scalar quantizers, denoted $Q_1(t, f)$ and $Q_2(t, f)$, with respective resolutions $R_1(t, f) < R_2(t, f)$, and a common scale factor $A(t, f)$ (hence $Q_1(t, f)$ is a sub-grid of $Q_2(t, f)$). The quantizers are defined for every f frequency bin, and they are updated every $t \times L$ samples of signal, where L is the length of a block of signal (typically every 2 seconds), to take into account the variability of the coefficient dynamics across frequency and time. The watermark on each MDCT coefficient corresponds to the modification of its amplitude from its quantized level on the grid $Q_1(t, f)$ to an arbitrary "sub-level" on the grid $Q_2(t, f)$, this latter level being determined by the watermark content. Hence, the embedding capacity of each MDCT coefficient is given by:

$$C(t, f) = R_2(t, f) - R_1(t, f) \tag{2}$$

and the capacity of a $T \times F$ molecule M^{x}_{pq} is given by:

$$C_M(t, p) = T \times \sum_{i\in P} C(t, i) \tag{3}$$

To maximize $C(t, f)$, $R_1(t, f)$ has to be minimized and $R_2(t, f)$ has to be maximized. However, $R_1(t, f)$ has to be high enough so that the quantization of MDCT coefficients with $Q_1(t, f)$ remains inaudible. The scale factor of both quantizers, $A(t, f)$, is determined from the maximum MDCT coefficient $M_{max}(t, f)$, on each frequency bin and each L-block. Given this, a fixed resolution of $R_1(t, f) = 8$ bits is chosen since it has proved to be inaudible from extensive listening tests on a large set of music signals. As for $R_2(t, f)$, the upper limit is a consequence of the audio-CD format conversion of the watermarked mix signal. This conversion can be simulated by the addition of a uniform additive noise $b_{Q16}(n)$. In the MDCT domain, this noise is observed to be approximately white Gaussian with standard deviation σ_{16} independent of f (and t). We thus assume that the amplitude of the maximum deviation on MDCT coefficients caused by the time-domain quantization remains lower than $4\sigma_{16}$. For the watermark to be preserved after audio-CD formatting, the quantization step $\Delta_{R_2}(f) = 2A(t, f)/2^{R_2(t,f)}$ has to verify:

$$4\sigma_{16} < \frac{\Delta_{R2}(f)}{2} \tag{4}$$

Hence the following condition on $R_2(t, f)$:

$$2^{R_2(t,f)} < \frac{A(t, f)}{4\,\sigma_{16}} \tag{5}$$

As shown in the result section, this condition generally enables significant embedding capacity, since the power of the 16-bits quantization noise is low.

Contrary to coding algorithms, in the proposed source separation method the quantizer parameters are not transmitted to the decoder, although they are necessary for the extraction of the source descriptors. They must be retrieved from the watermarked mix signal, despite of the modification of the MDCT coefficients by the watermark. Actually, only $A(t, f)$ has to be retrieved since R_1 is fixed and $R_2(t, f)$ is specified by (5). Since $A(t, f)$ is determined from $M_{max}(t, f)$, a solution is to quantize the resulting value using a 6-bits quantizer $Q_A(f)$, which is i) known at both the encoder and the decoder, ii) independent of time, and iii) insensitive to the watermarking process (this point requires some simple additional processing that will not be detailed here).

2.3. Source signals descriptors and their use for separation

The descriptors of the molecules of source signals depend on the embedding capacity of the corresponding molecule (*i.e.* the molecule with same coordinates in the TF plane) of the host mix signal. In the case of a poor capacity, the energy ratio between a molecule of a source signal s_k and the corresponding molecule of the mix x is used:

[1] For the same reason, the proposed method is not suitable for compressed signals.

$$E_{s_k/x}(p,q) = \frac{\sum\limits_{(i,j)\in\{P\times Q\}} |m_{ij}^{s_k}|^2}{\sum\limits_{(i,j)\in\{P\times Q\}} |m_{ij}^{x}|^2} \tag{6}$$

This ratio is quantized to $\check{E}_{s_k/x}(p,q)$ using a scalar quantizer, and the resulting index is embedded. The source signal molecule is reconstructed at the decoder by the corresponding mix molecule weighted by $\sqrt{\check{E}_{s_k/x}(p,q)}$. This is done for each molecule of each source. Therefore, the separation is based here on molecular energy segregation.

In the case several sources overlap, the shape of the mix molecule can be quite different from the shape of the source molecule it is supposed to reconstruct. For this reason, if $C_M(t,p)$ is large enough, additional information describing the structure of source molecules is embedded. Hence, molecule prototypes are used as *shape* descriptors. Codebooks of (matrix) prototype shapes are used to strongly reduce the coding cost of this descriptor (compared to individual MDCT coefficient coding). These codebooks are designed for each molecular frequency bin p using a Linde-Buzo-Gray-based algorithm [9] applied on a large database of speech/music signal molecules (note that these codebooks can be adapted to represent a given instrument, or a given type of voice). The shape descriptor of a source molecule is the closest prototype shape in a codebook, according to the Euclidean distance. As in vector quantization techniques, the codebooks are assumed to be known at the decoder, and the index of the prototype in the codebook is embedded as shape information. Note that to increase the codebooks efficiency in term of quality/coding-cost ratio, molecules are normalized before coding. A molecule $M_{pq}^{s_k}$ with mean (across the MDCT coefficients of the molecule) $\mu_{pq}^{s_k}$ and standard deviation (idem) $\sigma_{pq}^{s_k}$ is normalized by:

$$N_{pq}^{s_k} = \frac{M_{pq}^{s_k} - \mu_{pq}^{k}}{\sigma_{pq}^{s_k}} \tag{7}$$

Additional descriptors $\mu_{pq}^{s_k}$ and $\sigma_{pq}^{s_k}$ are encoded separately, as in, e.g., [10], using scalar quantizers, to provide $\check{\mu}_{pq}^{s_k}$ and $\check{\sigma}_{pq}^{s_k}$. Note that the parameters of the additional scalar quantizers are themselves quantized and transmitted via watermarking (as for $A(t,f)$ in Section 2.2). The cost of this additional embedding is assumed to be very low in comparison to the cost of descriptors, since those parameters are encoded only once in a given L-block of signal.

At the decoder, if M_{i_q} denotes the molecule of the shape codebook $D(p)$ closest to $M_{pq}^{s_k}$, $M_{pq}^{s_k}$ is estimated by:

$$\hat{M}_{pq}^{s_k} = \check{\sigma}_{pq}^{s_k} \times M_{i_q} + \check{\mu}_{pq}^{s_k} \tag{8}$$

Finally, the source signals are reconstructed from the estimated source molecules by applying inverse MDCT.

2.4. Bit allocation

Distributing the embedding resource of (3) among the different descriptors is a complex optimization problem, since it depends on the number of sources to separate, their nature (an instrument and a speech signal do not need the same amount of data to be accurately described), the nature of the descriptors, and their coding precision. Moreover, limits for the size of codebooks must be taken into account, since too small a codebook cannot represent efficiently the shape of a molecule, while too large a codebook would considerably increase the coding/decoding computational cost. In the present study, a series of bit allocation tables were determined empirically for different separation configurations, and validated by listening tests. The bit allocation tables need to be known both at the encoder and the decoder. Table 1 is an example designed for the separation of 2 speech signals. Note that the streaming of embedded data among the MDCT coefficients of a molecule is a trivial task and is not detailed here.

$C_M/2$	8 to 13	14	15	16	17	18	19	20	21	22	23	24	25
Shape	0	6	7	7	7	7	8	8	8	9	9	10	10
Gain	$C_M/2$	4	4	5	6	6	6	7	7	7	8	8	8
Mean	0	4	4	4	4	5	5	5	6	6	6	6	7

Table 1: *Bit allocation table (per source) for the separation of 2 speech sources with molecules of size* 2×4

3. EXPERIMENTS AND RESULTS

Tests have been processed with both speech and music signals with speech+speech mixtures, and speech+instruments mixtures, with a number of sources varying from 2 to 4. Instruments are a bass guitar, a classic guitar and a flute. All signals are at the audio-CD format (linear 16-bits, 44.1kHz sampling). Speech signals are from 3 male and 3 female English speakers. In the present paper, results are presented for speech separation among several speech signals or music signals. Separation of instruments will be developed in further research.

The quality of separated sources has been assessed by both listening tests and ISNR measures, as defined in [11] (ISNR is the difference (Improvement) of log-Signal-to-Noise ratios between input and output of the separation process, where Signal is the source signal to be separated, at the input Noise is the sum of all other signals in the mixture, and at the output Noise is the difference between original and estimated source signal).

3.1. Molecule size and embedding capacity

Concerning the molecule size, a trade-off had to be found between a sufficient capacity that enables to encode descriptors properly, and a correct resolution in the TF plane molecules. Figure 2 shows the separation results obtained for 5 different sizes of molecule, containing from 8 to 16 MDCT coefficients. For this experiment, the descriptors of the sources were not quantized (only the "separation power" of the molecular decomposition is tested). A 2×4 molecule size appears to be the best trade-off. Results are slightly lower for a 4×2 molecule, which confirms the importance of spectral dynamics for audio signals. Figure 3 gives the embedding capacity for each MDCT coefficient as a function of frequency bin, and the resulting capacity of a 2×4 molecule in the case of a mixture of 2 speech signals. The energy concentration of speech and music signals in low frequencies results in a large embedding capacity, up to 60 bits per molecule. In high frequencies, fewer bits are available to insert the watermark, but in this region the human ear is less sensitive, and source signals can be represented with a lower accuracy. Thus, the ratio between the information to embed and the embedding capacity can be satisfied along all frequency bins.

3.2. The complete separation system

In this section we report preliminary separation results obtained with the complete system. Note that it was verified that watermarking MDCT coefficients with $Q_2(t,f)$ does not impair the audio quality of the mix signal (this is not surprising since it was the case with

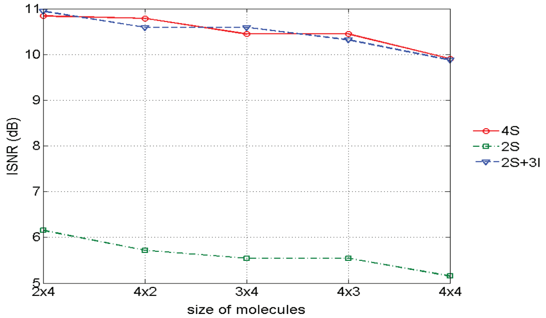
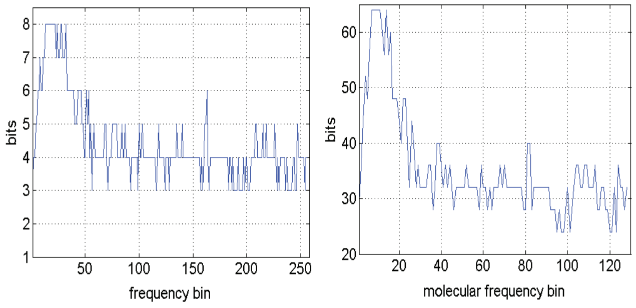


Fig. 2: Separation performance vs. molecule size.



(a) MDCT coefficients

(b) Molecules

Fig. 3: Embedding capacity (in bits) as a function of frequency for MDCT coefficients (a) and molecules (b).

$Q_1(t, f)$ which is coarser than $Q_2(t, f)$). Figure 4 provides the results obtained for the two configurations of descriptors (energy ratio (ER) *vs.* shape + mean + standard deviation (SMSD)), and for the separation of 2 speech signals in a 4 speakers mixing (4S) and in a 2 speakers + 3 instruments mixing (2S+3I). In each case, six different test signals were used, with a duration of approximately 3s. The bit allocation of Table 1 was used for those tests.

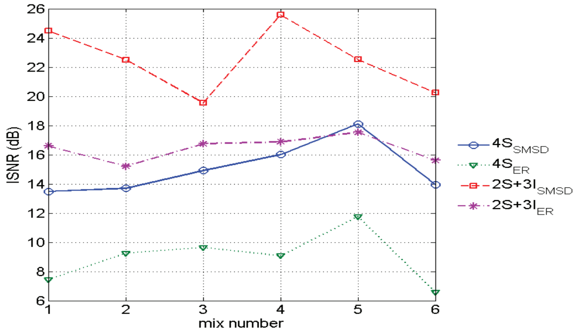


Fig. 4: Separation performance for ER and SMSD configurations.

This figure exhibits ISNR scores from about 10dB up to 25dB depending on the configuration, hence a good separation of the speech source signals. Listening tests performed in the ER configuration reveal remaining interferences on the separated sources. Figure 4 proves the improvement obtained by using shape codebooks, with an average ISNR increase across configurations of about 7dB. This gain results in a very significant improvement in the quality of separated signals.

4. CONCLUSION

The separation approach described in this paper does not belong to classical source separation methods. Contrary to the BSS framework, in ISS, source signals are available before the mix is processed, and specific (but important) applications such as audio-CD "active-listening" are targeted. In the present study, promising preliminary results on speech signals separation from a single-channel mixture have been reported.

Future work will mainly focus on music instruments separation. The use of shape codebooks for each kind of instrument is expected to provide efficient source representation and thus separation. Also, the use of refined TF decomposition algorithm such as Molecular Matching Pursuit [5] [6] and/or multi-resolution decomposition will also be considered to better take benefit of the audio signals sparsity, and provide a more accurate separation. Finally, the ISS framework can be extended to the multi-channel source separation framework, where watermarking can be combined with more powerful source separation techniques.

5. REFERENCES

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

[2] C. Jutten and J. Herault, "Blind separation of sources, part 1. an adaptative algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[3] D.F. Rosenthal and H.G. Okuno, Eds., *Computational auditory scene analysis*, Lawrence Erlbaum Associates, Inc. Mahwah, NJ, USA, 1998.

[4] J.P. Princen and A.B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. 64, no. 5, pp. 1153–1161, 1986.

[5] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 14, no. 5, 2006.

[6] M. Parvaix, S. Krishnan, and C. Ioana, "An audio watermarking method based on molecular matching pursuit," in *IEEE ICASSP*, 2008, pp. 1721–1724.

[7] B. Chen and G. Wornell, "Quantization index modulation : a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Information Theory*, vol. 47, pp. 1423–1443, 2001.

[8] K. Brandenburg and M. Bosi, "Overview of mpeg audio: Current and future standards for low bit-rate audio coding," *Journal of the Audio Engineering Society*, vol. 45, no. 1, pp. 4–21, 1997.

[9] Y. Linde, A. Buzo, and R. M. Gray, "Algorithm for vector quantizer design," *Trans. IEEE Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[10] K. L. Oehler and R. M. Gray, "Mean-gain-shape vector quantization," in *IEEE ICASSP*, 1993.

[11] 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.