Effects of Audio Coding on ICA Performance: an Experimental Study

Matthieu Puigt*†, Emmanuel Vincent‡, Yannick Deville§, Anthony Griffin†, and Athanasios Mouchtaris†¶
*LISIC, ULCO, Université Lille Nord de France, Calais, France, FR-62228
email: matthieu.puigt@lisic.univ-littoral.fr
†FORTH-ICS, Heraklion, Crete, Greece, GR-70013
email: {agriffin,mouchtar}@ics.forth.gr
‡ Inria, Villers-lès-Nancy, France, FR-54600
email: emmanuel.vincent@inria.fr
§IRAP, Université de Toulouse, CNRS, Toulouse, France, FR-31400
email: yannick.deville@irap.omp.eu
¶University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409

***Abstract*—In this paper, we study the influence of lossy audio coding on the performance of Independent Component Analysis (ICA). In particular, we derive two compression scenarios from practical implementations. In the first case, we consider the situation when the sources are independently compressed, decompressed and then mixed. In the second, we consider the situation when mixtures of sources are jointly compressed. We experimentally show that the modification of the spectro-temporal diversity due to compression has almost no effect on the performance of ICA methods. We also show that the two tested stereo encoding strategies have a major effect on the performance of ICA, especially when the mixed signals are compressed at low bit rates. As these strategies have been extended to audio systems involving much more than two channels, our work suggests that ICA will not be able to successfully process the high number of observations provided by modern immersive audio systems.**

I. Introduction

We live in a digitized world: we are now accustomed to watching videos, viewing images and listening to music using computers or mobile devices. As the storage of such content (or its transmission through a network) may be costly, several lossy encoding formats allowing data compression with an acceptable rendering from a human point of view have been proposed, e.g., MPEG-x for videos/music [1] or JPEG for images [2]. Until a few decades ago, people were working with RAW data before compressing them in a post-processing stage. This is no longer the case: we are used to directly retouching JPEG images obtained with digital cameras (which do not provide RAW data anymore in most cases), and we listen to our songs as MP3s instead of the original WAV files.

Blind Source Separation (BSS), and in particular Independent Component Analysis (ICA), have proved to be great tools for image and audio processing [3], and the limits of the independence assumption in natural images [4] and in audio signals [5] are well known. Interestingly, while the use of ICA for compressing data has been investigated for image [6] or audio [7] signals, to the best of the authors' knowledge, the influence of data compression on the performance of ICA has never been studied. Here we propose to explore this relationship for audio signals, in two mixing scenarios that we define below.

Scenario 1: We face an "uncompressed mixture of compressed signals" when the sources are first individually compressed, then decompressed, and eventually mixed together.

Such a situation is encountered in VoIP: in a videoconference, each site sends compressed audio information [8] and then receives a mixture of the decompressed audio sources provided by the other users. This scenario also occurs in multi-track audio formats [9], where the original tracks (i.e., the sources in a BSS problem) are individually encoded in a classical audio format (including lossy encoding formats like MP3), and are then mixed by the user who can export such a mix as a stereo WAV file.

Scenario 2: We face a "compressed mixture of uncompressed signals" when the sources are first mixed together and the resulting mixture is then compressed.

This is the classical situation encountered in audio processing when people listen to music stored in their MP3 players: a mixture is first created by recording several sources simultaneously or mixing them via a software and this mixture is then compressed. In particular—as with many audio coding approaches—when a WAV file is converted to MP3, some harmonics are filtered out and some quantization noise is added. Moreover, stereo encoding might have an influence on the spatial diversity contained in the recordings.

In this paper, we study the above mixing scenarios in both the time and the time-frequency (TF) domains, following the framework of our previous work in [5]. To this end, we focus on the effects of MP3 coding, as it is very commonly used and involves encoding strategies [1] shared by many other compression techniques. We explore its influence on dependency measures and/or on the performance of ICA. In the remainder of this paper, we investigate the effects of the first and second compression scenarios on performance in Sections II and III respectively.

This work was funded partly by the Marie Curie IAPP "AVID MODE" grant within the European Commission's FP7, and partly by "Axe transverse analyse et traitement de données" of Observatoire Midi-Pyrénées (Toulouse, France).

II. Uncompressed mixtures of compressed signals

In this section, we consider the first mixing scenario. As explained in the introduction, MP3 coding is one of the most popular methods for compressing audio data. As with many encoding approaches, it consists of [1]: computing a time-frequency transform of the signal; then analyzing each temporal frame by a psychoacoustic model which indicates the masking curve, i.e., the thresholds per subband below which the quantization noise will be inaudible due to masking effects; and finally, applying a quantization process which ensures that quantization noise is shaped in the frequency domain so as to be below the masking curve and thus will remain inaudible. In this section, we explore what effect this spectro-temporal compression has on the performance of ICA methods, in both the time and time-frequency domains.

A. Analysis in the time domain

Many ICA techniques achieve source separation by minimizing some dependency measure between the estimated source signals. Previous work [5] studied the influence of the nature of audio signals on the performance of ICA methods as a function of the excerpt length. Stationary and nonstationary dependency measures between sources were estimated and validated by some ICA simulations. Moreover, the nonstationary measure called Gaussian mutual information (GMI) was shown to be better suited to audio signals. In this paper, we stick to the GMI, defined as

$$\mathcal{GI}\{s_1, \ldots, s_N\} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{2} \log \frac{\det \operatorname{diag} \widehat{\mathbf{R}}_s(q)}{\det \widehat{\mathbf{R}}_s(q)}, \qquad (1)$$

where Q is the number of disjoint time frames over which the source covariance matrices $\widehat{\mathbf{R}}_s(q)$ ($q \in \{1 \ldots Q\}$)—whose (i,j)-th element is the cross-covariance between sources $s_i(q)$ and $s_j(q)$—are computed. Note that we do not need to normalize the signals, since this measure does not depend on their scales. The separation performance is related to the value of the GMI over the true source signals, which was shown to be near zero for long excerpt duration and much higher than zero for the shortest excerpts [5].

In this section, as we independently encode each source, we investigate what effect the spectro-temporal compression has on the GMI in the time domain. We consider the audio BSS dataset [10], which consists of thirty pairs of speech sources and thirty pairs of music sources, sampled at 22.05 kHz. These signals are collected from English audio books read by different speakers and from synchronized multitrack recordings, respectively. They are also resampled at 8 kHz and 16 kHz, respectively, in successive tests, in order to keep the same sampling frequency as the signals compressed at the lowest tested bit rates. They are then independently encoded as mono MP3 files by the LAME encoder[1] with different bit rates from 8 to 160 kbps. The above 540 pairs of signals are then split into disjoint excerpts of equal durations, from 2^7 samples to 2^{16} samples. The GMI is computed for each of these excerpts where we set the number Q of frames in Eq. (1) to $Q = 8$. Figure 1 shows the average GMI, estimated over all excerpts and all sources for the different bit rates of MP3 files and for WAV files. It shows that compression has very little effect on the GMI, with slightly more influence on music signals than on speech. Compressed music signals exhibit slightly lower dependency measures than RAW ones. The standard deviation of these measures—not shown for space considerations—is similarly not affected by compression.

[1] See http://lame.sourceforge.net/.

B. Analysis in the time-frequency domain

Many ICA methods work in the time-frequency domain, e.g., for convolutive mixtures. It then makes sense to study the behavior of the GMI in the time-frequency domain. Moreover, because of the time-frequency processing underlying MP3 conversion, we might have a bigger influence of the encoding bit rate in this domain than in the time domain. We thus repeated the experiment proposed in [5]: we still used the signals [10], processed as in the previous section, i.e., also resampled at 8 kHz and 16 kHz and converted to MP3, and we computed their short-time Fourier transform (STFT). In these tests, the length of the STFT window function geometrically increases from 2^7 to 2^{13} samples. For each window length, for each pair of source signals and for each frequency bin, we compute the GMI. We then derive the mean values of the GMI—not shown for space considerations—over all frequency bins, with respect to the length of the windowing function, computed with WAV and MP3 files for different bit rates. Like in the time domain experiment, the compression has a very limited effect on the GMI which is slightly higher for music signals than for speech signals. Again, the MP3 music signals are slightly less dependent than the WAV ones.

C. Discussion

It is known that speech signals are sparser than music [11]. For example, the time-frequency supports of several arbitrary speech signals are approximately disjoint while those of music signals are usually not [11]. This means that speech signals need (much) less bits to be encoded than music, for the same perceived quality. Interestingly, our tests showed that the increase of sparsity due to compression slightly decreased the source dependency measures of the less sparse signals we considered, i.e. music, while it was not the case for speech signals. This implies that the connections between source independence and joint sparsity are probably more subtle than what is already known in sparse coding [4], as it was recently shown for fMRI [12].

Our results also suggest that compressing music source signals should improve their separation, not only with ICA—since sparser signals get more points around zero and are thus more super-Gaussian—but also with the other classes of BSS methods assuming the sources to be sparse, i.e., Sparse Component Analysis (SCA) and Non-negative Matrix Factorization (NMF). In particular, we expect to improve the separation in the underdetermined case, i.e., when the number of sources in stereo recordings is higher than two. For space considerations, such an investigation is out of the scope of this paper and is left for future work.

III. Compressed mixtures of uncompressed signals

We briefly described in Section II how a mono signal is compressed when encoded with an MP3 encoder. When the

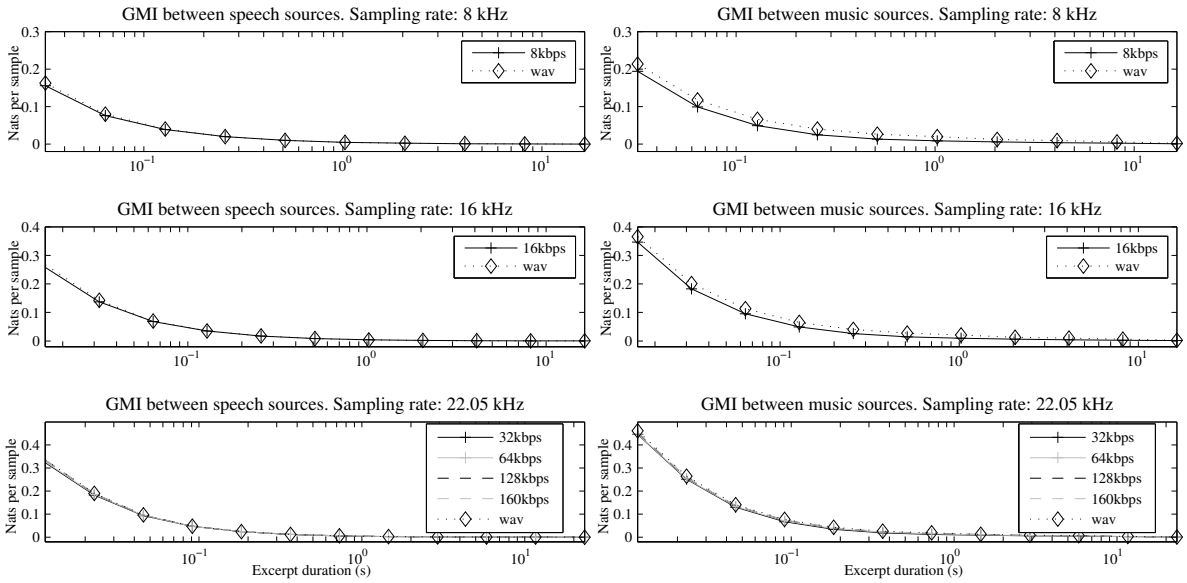

Figure 1. Mean GMI obtained in the framework of the first scenario, with MP3 and WAV files in the time domain.

signal is stereo, LAME automatically switches between two strategies, named *modes*, to encode it: stereo, and joint-stereo.

The stereo mode consists of independently compressing each channel, while negotiating the bit demand between both channels.

The joint-stereo mode exploits the interchannel dependencies to reduce the overall bit rate [1]. It reformulates the observations as "middle" (the sum of the two channels) and "side" information (the difference between them), and quantizes both of them while providing a higher bit rate to the middle information. As a consequence, the source positions vary over time and frequency. They may even become equal for several sources in some frequency bins, such that the mixing matrix becomes non-invertible and the mixture cannot be separated by BSS anymore. In this mode, a more aggressive option, named *intensity stereo*—not provided in LAME but present in other encoders—consists of coding the upper-frequency subbands of the middle information *only*. The decoder then reconstructs the left and right channels by using *only* the middle channel and independent left and right scale factors[2]. As a consequence, the spectral shape of each channel will be the same for these subbands, up to a scale factor, and the mixing matrix becomes non-invertible in all upper frequency bins.

The scope of this section is to provide an experimental measure of the influence of the varying source image position—due to both encoding strategies—on the global ICA performance. In particular, LAME proposes an option to force the stereo mode only, so that we can finally apply two encoding strategies: the stereo mode only or the default setting which switches between the two above modes. We thus test both strategies below.

[2]The intensity mode is also widely used in multi-channel coding, i.e., the configuration where the number of loudspeakers is strictly greater than two.

A. Analysis in the time domain

As explained above, we now investigate the second compression scenario with respect to the choice of the two-channel audio encoding scheme. Let us stress again that, by default, LAME switches between the above modes, whose choice is based on the similiarities between the two channels. Moreover, even if the LAME encoder does not provide the intensity option in the joint-stereo mode, when the available bit rate is low (e.g., 8 kbps), LAME may not put any bits in the side channel, thus providing a kind of intensity mode. As a consequence, the influence of compression depends on the mixing parameters to be estimated.

We consider the pairs $s(t)$ of signals from the audio dataset [10], that we mix with symmetrical matrices defined as

$$A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \quad \text{with } \alpha = 0.1,\ 0.5, \text{ and } 0.9. \tag{2}$$

The resulting observations $x(t)$ then read

$$x(t) = A\, s(t). \tag{3}$$

In addition to the original mixtures, we also resample the above pairs of signals at 8 kHz and 16 kHz, respectively, in successive tests. These mixtures are passed through the LAME encoder, using the stereo or the default modes, with the same bit rates as in Section II. We use the same experimental protocol as before. However—and unlike in the previous section—we cannot directly estimate the GMI or the SIR because the definition of the "real" source images—i.e., the contribution of the sources in the compressed/decompressed mixtures—is not obvious. Indeed, quantization noise partly contains some source-dependent noise which should be associated with the corresponding sources and some additional noise components. There would be many ways to define source images, depending how the above noise is taken into consideration. Moreover, as joint-stereo encoding may destroy the spatial diversity of the sources in the mixtures for some subbands of some frames,

the definition of time-varying source position image is even more complex. Here we propose to overcome this limitation by using the Mixing Error Ratio (MER) [14]. For the i-th column of A, we rewrite the estimated mixing column $\widehat{a_i}$ as

$$\widehat{a_i} = a_i^{\text{coll}} + a_i^{\text{orth}}, \tag{4}$$

where a_i^{coll} and a_i^{orth} are respectively the vectors which are collinear and orthogonal to the *true* vector a_i. The MER associated with this column, denoted MER_i hereafter, then reads

$$\text{MER}_i = 20 \cdot \log \left(\left\| a_i^{\text{coll}} \right\| / \left\| a_i^{\text{orth}} \right\| \right). \tag{5}$$

The global MER is computed as the mean over i of the above MER_i. The variations of the MER with respect to the quantization level thus give insight into the influence of the stereo and joint-stereo modes on the BSS performance.

We thus apply the ICA method [13] to the generated 3240 pairs of mixed signals and estimate the MERs. Figures 2 and 3 provide mean MERs in the time-domain experiments with the highest tested bit rates, for the stereo and default modes, respectively. The plots with lower bit rates are not represented for space considerations. When $\alpha = 0.9$, the performance obtained in the stereo mode is always very low. This may be due to the fact that each channel is encoded independently. From a BSS point of view, this implies that the observations are no longer seen as linear instantaneous mixtures of sources. This phenomenon is less visible for lower values of α. Indeed, prior to the compression step, observations can then be (roughly) modeled as the contribution of their main source with some additive "noise" (which consists of the interfering source which has a low energy because of the low values of α). The resulting observations are thus sparser [11] for low values of α than when $\alpha = 0.9$. During the compression stage, the content which is filtered out is more likely to be negligible, thus barely changing the nature of the mixtures in observations. The default mode provides higher performance (but still 5 to 10 dB lower than with WAV files for the longest durations). When $\alpha = 0.9$, we notice that the MERs obtained at 8 kbps—not plotted here for space considerations—are much lower than those obtained at 16 kbps, while they were quite similar for lower values of α (and even very close when $\alpha = 0.1$). Actually, at the lowest bit rate, the joint-stereo mode does not provide any bits to the side channel in some frames, resulting in a complete loss of spatial diversity. For higher bit rates, we noticed that observations may be locally seen as mixtures of the middle and side channels instead of mixtures of the sources, which decreased the mixing matrix estimation accuracy, and thus the performance. This result is consistent with the above analysis. Lastly, when $\alpha = 0.1$, both encoding strategies provide almost the same performance, which was expected as the default setting selects the joint-stereo mode when both observations are almost similar, which is not likely the case for this value of α.

B. Analysis in the time-frequency domain

We repeated the same experiment as in Section III-A but in the time-frequency domain, following the framework described in Section II-B. Let us stress again that, as the main MP3 conversion stages are performed in the time-frequency domain, we might expect the bit rate values to have more influence on the BSS performance than in the time domain.

Figure 4 shows the variations of the mean MERs[3] obtained with signals encoded with the default setting. Even if most of the behaviours are consistent with the above study in the time domain, we noticed some interesting effects. For example, when $\alpha = 0.9$—i.e., when the joint-stereo encoding is more likely to be applied in the default setting—for the lowest bit rates, the MERs slightly increase with the STFT window size (this phenomenon is visible in Fig. 4 at 64 kbps, for example). This phenomenon seems to be strange as it was shown in our previous work [5]—and also on the same figure with WAV signals—that the contrary should occur: the larger the STFT window size, the lower the number of samples per frequency bins, thus making the statistics roughly estimated. Actually, the MERs increase when the STFT windows size is higher than 1024 samples, which almost corresponds to the frame size used by LAME to encode the signals. The observed phenomenon may be explained as follows. When the STFT window size in the BSS method is lower than 1024—and as discussed above—several frequency bins are expressed as linear combinations of the middle and side channels only, thus changing the nature of the mixing process and making the ICA method less accurately estimate A. On the contrary, when this size is higher than 1024, the spectro-temporal observed points contain both the combinations of middle and side channels—due to the joint-stereo mode—and the combinations of the left and right channels—due to the stereo mode. The source image position is then somewhat averaged in these bins. The overall quality of the samples is thus higher than for the lower STFT window sizes and compensates for the loss of performance that would be due to the lack of temporal data per frequency bin.

When the signals are encoded with the stereo mode, as depicted in Fig. 5, the MERs are the same as those with the default setting when $\alpha = 0.1$. For higher values of α, as in the time domain experiment, the MERs obtained with the stereo mode are (much) lower than those obtained with the default mode. This shows again the importance of the choice of the encoding strategy.

C. Discussion

We now explore which implications the above results might have for other BSS problems.

First of all, let us note that many stereo audio recordings—e.g., the songs recorded in studio—contain more than two sources, providing the so-called underdetermined BSS configuration. This implies that the spectro-temporal content to be compressed is richer, i.e., less sparse. As a consequence, the effects due to compression we noticed in this section should be stronger with more sources.

In almost all the recordings of songs performed in a studio, the main vocal has a centered source position, while the other instruments have different positions. The higher the number of sources, the higher the similarities between channels, thus increasing the chances to select the joint-stereo

[3]Here the MER_i is computed for each frequency bin and then averaged along the frequencies. All the source permutations along the frequencies are tested and only the best MER is kept.

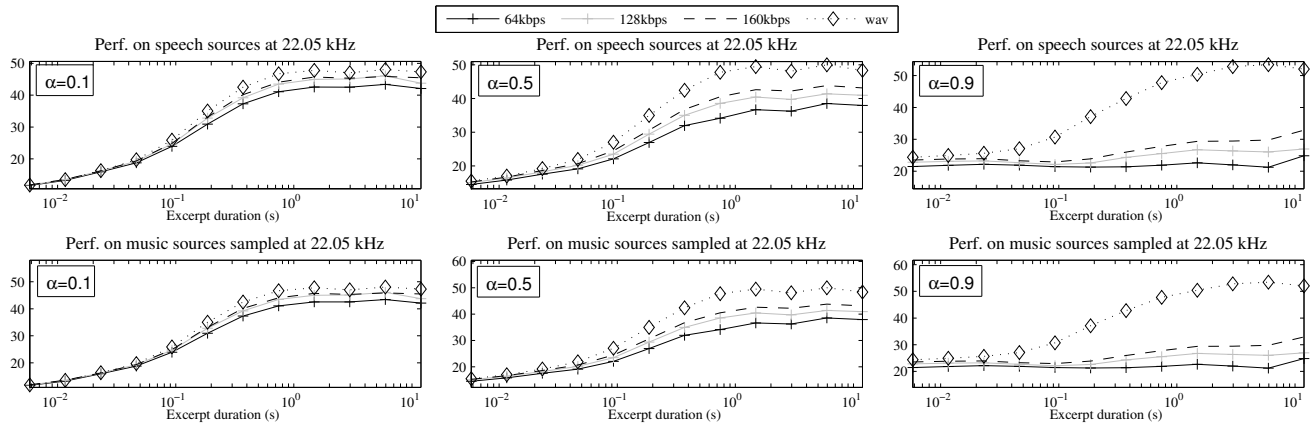

Figure 2. Mean mixing error ratio obtained in the framework of the second scenario, with the Pham-Cardoso method applied in the time domain, vs the excerpt duration, the mixture rate α, and the compression bit rate (in stereo mode).

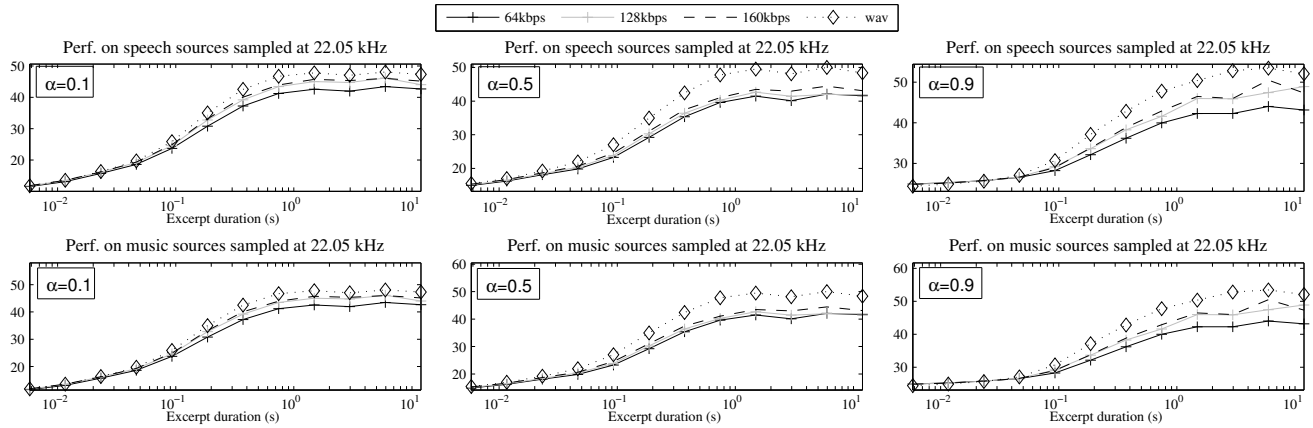

Figure 3. Mean mixing error ratio obtained in the framework of the second scenario, with the Pham-Cardoso method applied in the time domain, vs the excerpt duration, the mixture rate α, and the compression bit rate (in default mode).

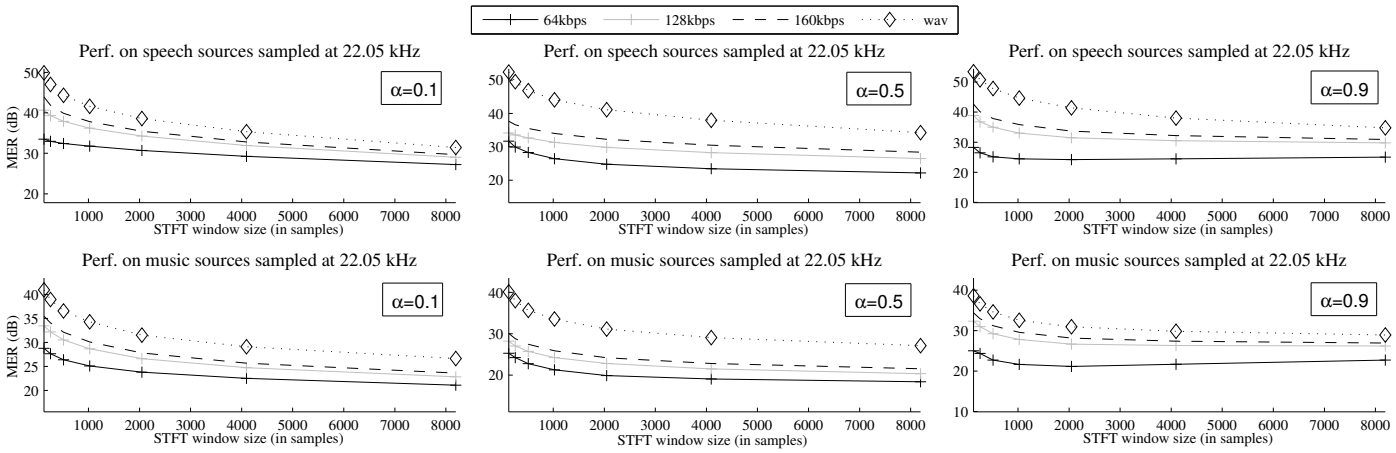

Figure 4. Mean mixing error ratio obtained in the framework of the second scenario, with the Pham-Cardoso method applied in the time-frequency domain, vs the STFT window size, the mixture rate α, and the compression bit rate (in default mode).

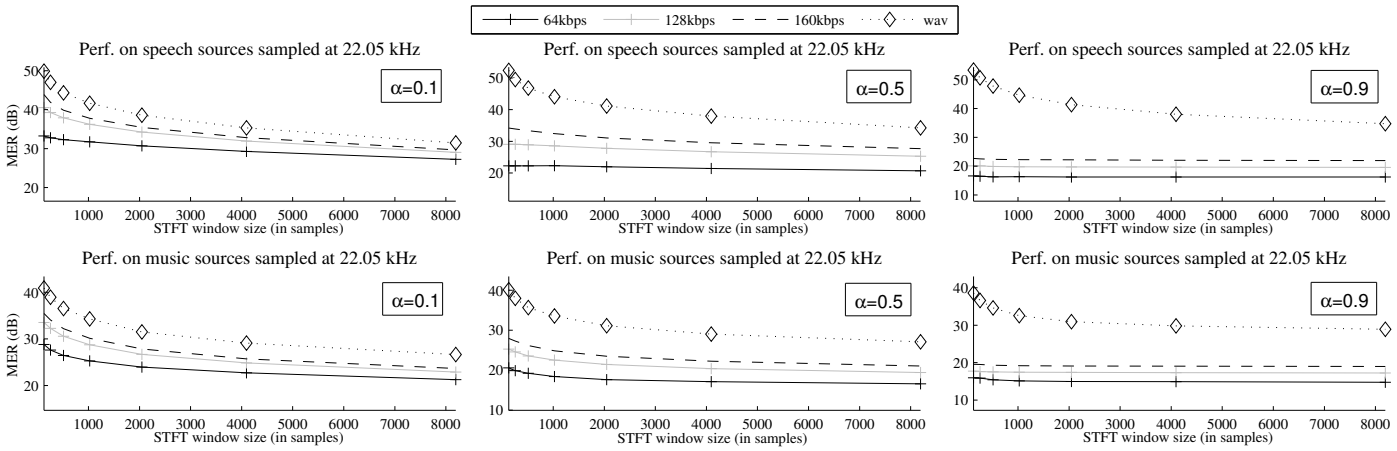

Figure 5. Mean mixing error ratio obtained in the framework of the second scenario, with the Pham-Cardoso method applied in the time-frequency domain, vs the STFT window size, the mixture rate α, and the compression bit rate (in stereo mode).

mode. Moreover, as the content to be compressed should not be really sparse, there should be extremely few bits kept for the side information. We already discussed the effects of this intensity-like encoding. The resulting observations might not be invertible for all frequency bins and time frames. Such an issue should be met with all the families of BSS methods, i.e., ICA, SCA, and NMF, and might be met with other encoding methods which process more than two channels. This is for example the case of *immersive audio systems*, which consist of synthesizing a 3D-audio environment with tens of loudspeakers. As the number of observations would be higher or equal to the number of sources, from a naive point of view, one could expect ICA to process signals from such systems with accuracy. However, for storage reasons, these systems process audio content which is compressed with extensions of intensity-coding-based techniques and our results show that it is reasonable to presume a low ICA performance.

IV. Conclusion

In this paper, we studied the influence of audio coding on the performance of Independent Component Analysis (ICA). In particular, we derived two compression scenarios from practical implementations: in the first case, we considered the situation when sources were independently compressed, decompressed and eventually mixed. In the second one, we considered the situation when mixtures of sources were jointly compressed. We experimentally showed that the modification of the spectro-temporal diversity due to mono compression had almost no effect on the performance of ICA methods, whereas that of the spatial diversity due to stereo compression had a large effect at low bit rates. Moreover, we showed the influence of the stereo encoding strategy on the quality of separation. This shows the need for new-generation BSS approaches for encoded mixtures. In particular, our results suggest that standard ICA—and more generally standard BSS—may not get benefits from the higher number of observations contained in multichannel systems, as they are encoded with extensions of MP3 compression schemes. This implies the need for new-generation source separation methods which are robust to the compression strategy. Recently, non-blind techniques such as informed source separation [15] could address such a problem. However, blindly doing the same work is still an open challenge.

References

[1] P. Noll, *MPEG digital audio coding standards*. CRC Press LLC., 2000

[2] G.K. Wallace, "The JPEG still picture compression standard," Commun. ACM, Vol. 34, No. 4, pp. 30–44, 1991

[3] P. Comon and C. Jutten, *Handbook of blind source separation. Independent component analysis and applications*. Academic press, 2010

[4] A. Hyvärinen, J. Hurri, E. Oja, *Natural image statistics – a probabilistic approach to early computational vision*. Springer-Verlag, 2009

[5] M. Puigt, E. Vincent, and Y. Deville, "Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources," in *Proc. of ICA*, Vol. 5441, pp. 613–620, 2009

[6] M. Narozny, M. Barret, and D.T. Pham, "ICA based algorithms for computing optimal 1-D linear block transforms in variable high-rate source coding," Signal Processing, Vol. 88, No. 2, pp. 268–283, 2008

[7] A. Ben-Shalom, S. Dubnov, and M. Werman, "Improved low bitrate audio compression using ICA without psychoacoustic modeling," in *Proc. of ICASSP*, pp. V.461–V.464, 2003

[8] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," IEEE Trans. on Speech and Audio Proc., Vol. 13, No. 5, pp. 787–798, 2005

[9] F. Gallot, O. Lagadec, M. Desainte-Catherine, and S. Marchand, "Iklax: a new musical audio format for interactive music," in *Proc. of ICMC*, pp. 85–88, 2008

[10] E. Vincent, R. Gribonval, and M.D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," Signal Processing, Vol. 87, No. 8, pp. 1933–1950, 2007

[11] D. Giannoulis, D. Barchiesi, A. Klapuri, and M.D. Plumbley, "On the disjointess of sources in music using different time-frequency representations," in *Proc. of WASPAA*, pp 261–264, 2011

[12] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'Ardenne, W. Richter, J.D. Cohen, and J. Haxby, "Independent component analysis for brain fMRI does not select for independence", PNAS, Vol. 106, pp. 10415–10422, 2009

[13] D.T. Pham, J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," IEEE Trans. on Signal Processing, Vol. 49, No. 9, pp. 1837–1848, 2001

[14] E. Vincent, S. Araki, and P. Bofill,"The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation," in *Proc. of ICA*, Vol. 5441, pp. 734–741, 2009

[15] L. Girin and J. Pinel, "Informed audio source separation from compressed linear stereo mixtures," in *Proc. of AES Int. Conf.*, 2011