Validity of the Independence Assumption for the Separation of Instantaneous and Convolutive Mixtures of Speech and Music Sources

Matthieu Puigt[1], Emmanuel Vincent[2], and Yannick Deville[1]

[1] Laboratoire d'Astrophysique de Toulouse-Tarbes, Université de Toulouse, CNRS, 14 Av. E. Belin, 31400 Toulouse, France,
{mpuigt,ydeville}@ast.obs-mip.fr
[2] IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France,
emmanuel.vincent@irisa.fr

Abstract. In this paper, we study the validity of the assumption that speech source signals exhibit lower dependency and therefore better separability with Independent Component Analysis algorithms than music sources. In particular, we investigate some dependency measures in the temporal and the time-frequency domains, resp. in the framework of instantaneous and convolutive mixtures. Moreover, we test several ICA methods, based on the above dependency measures, on the same source signals. We experimentally show that speech and music sources tend to have the same mean behaviour for excerpt durations above 20 ms, but music signals provide more spread dependency measures and SIR values. Lastly, we experimentally show that Gaussian nonstationary mutual information is better suited to audio signals than mutual information.

1 Introduction

Independent Component Analysis (ICA) [1] is the most investigated class of methods to solve the Blind Source Separation (BSS) problem. Among the applications of BSS, audio processing is one of the major areas of interest. In this paper, we aim to study the behaviour of dependency measures when applied to speech or music source signals, with respect to the length of the signal recordings, for linear instantaneous and frequency-domain convolutive mixtures. Indeed, one generally assumes that, in the "cocktail party" problem, speech sources are independent and are thus separable thanks to ICA while, on the contrary, this is not the case for music sources, because musicians play in a coherent way, thus yielding dependent source signals [2].

A dependency measure, i.e. the mutual information, has been previously studied, only for speech sources, and only in the framework of linear instantaneous mixtures [3]. The authors show that speech source signals are independent (resp. have some dependencies) when source signal excerpts have long duration (resp. short duration). However, the chosen estimator has larger bias and variance than estimators developed more recently [4]. Moreover, [3] only provides mean values of the estimated mutual information and therefore omits the variance of

this measure. One also finds some papers in the literature, e.g. [5], which study the effects of the length of the time-frequency windows on the performance of frequency-domain convolutive methods. However, [5] only investigates the above effects as a function of the length of the impulse response of the mixing filters, for speech signals. It therefore does not show any possible difference of performance with respect to the nature of source signals. Moreover, the authors only measure the source correlation, which does not correspond to the dependency measure used in the tested ICA method.

As a consequence, in Sect. 2, we generalize [3] by computing the statistics (mean values and variance) of two dependency measures of speech *and* music sources (contrary to [3], we not only test the mutual information but also the Gaussian nonstationary mutual information), and by applying ICA algorithms to these signals. In Sect. 3, we extend the above procedure to the framework of frequency-domain convolutive ICA which was studied in [5] in more specific conditions. Conclusions are derived from this investigation in Sect. 4.

2 Independence and ICA Performance in Time Domain

Many ICA techniques achieve source separation by minimizing some dependency measure between the estimated source signals. In this paper, we study two classical measures, i.e. the zero-lag mutual information [1] and the Gaussian nonstationary zero-lag mutual information [6], resp. defined as

$$\mathcal{I}\{s_1,\ldots,s_N\} = -\mathbb{E}\left\{\log \frac{\mathbb{P}_{s_1}(s_1)\ldots\mathbb{P}_{s_N}(s_N)}{\mathbb{P}_{s_1,\ldots,s_N}(s_1,\ldots,s_N)}\right\} \tag{1}$$

and

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

where $\mathbb{E}\{.\}$ stands for expectation, $\mathbb{P}_{s_1,\ldots,s_N}$ and $\mathbb{P}_{s_i}$ $(i \in \{1\ldots N\})$ are resp. the joint and marginal probability density functions of the sources and Q is the number of disjoint time frames over which the source correlation matrices $\widehat{\mathbf{R}}_s(q)$ $(q \in \{1\ldots Q\})$ are computed. Note that we do not need to normalize the signals, since the above measures do not depend on their scales. The separation performance may be related to the value of these measures over the true source signals, which is assumed to be near zero. The validity of this assumption depends on the considered mixture. As explained in Sect. 1, while speakers at a "cocktail party" tend to speak freely without attention to distant speakers, musicians often play synchronous sounds at related frequency ratios as specified by the rules of music harmony [2], regardless of the recording duration. Therefore, for each of these types of signals, the dependency between such source signals should intuitively be the same for all durations. One may also expect it to be larger for music than for speech. We are going to study whether these intuitions are true.

We consider the audio BSS dataset [7], which consists of thirty pairs of speech sources and thirty pairs of music sources, sampled at 22.05 kHz. These signals

are resp. collected from English audio books read by different speakers and from synchronized multitrack recordings. All pairs of signals are then split into disjoint excerpts of equal durations, from 2^7 samples (5.7 ms) to 2^{18} samples (11.9 s). The mutual information is estimated via the software proposed[3] in [4] and we set the number Q of frames in (2) to $Q = 8$ for computing the Gaussian mutual information. The above dependency measures are computed for each above-defined excerpt.

Figure 1 shows the variations of the estimates of the Gaussian mutual information, for one pair of speech signals, with respect to the index of the excerpt and their time duration. The obtained values have a random behaviour over time. However, when the excerpt length increases, the highest measures of dependency significantly decrease. A similar behaviour has been observed for music sources and for mutual information dependency measures. Therefore, for the sake of brevity, it is not shown in Fig. 1. As a consequence, contrary to [3], we hereafter always study both the mean values and the associated standard deviations (that we consider as a spread measure) of these measures.

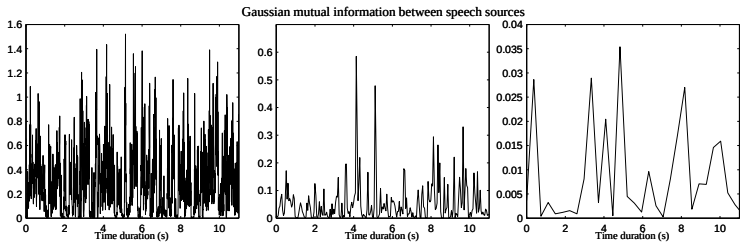

Fig. 1. Variations over time of the Gaussian mutual information between two speech signals in [7], computed for several lengths of test excerpt (*left:* 2^7 samples, *center:* 2^{10} samples, *right:* 2^{13} samples).

Figure 2 thus shows the mean values over all excerpts and all sources of the estimated mutual information and Gaussian mutual information and their above-defined spreads, vs excerpt length. In order to measure the variance of the estimators, this experiment was also conducted for thirty pairs of independent Gaussian white noise signals. These results show that both dependency measures are much larger for audio sources than for independent noise signals, regardless of their excerpt duration. This is partly due to the short-term periodicity of some speech and music sounds [3]. The results also show that both dependency measures span a larger range for music than for speech, but that they are similar for both types of sources on average, except for short durations, i.e. below 20 ms where there are higher for music. In this case, both mean measures of the dependency are high, thus showing that the independence assumption is not fulfilled

[3] Matlab code is accessible at http://www.klab.caltech.edu/~kraskov/MILCA/.

for speech nor for music, which is in agreement with [2,3]. However, they significantly decrease when the excerpt time duration increases. This phenomenon is observed for all pairs of sources, except for one pair of electronically-generated music sources whose mutual information keeps a high value for durations above 2 s due to repetitions of the same note samples over time. Similarly, keeping high dependencies might be observed in a civilized dialogue situation with one speaker being silent when the other speaks and vice-versa. Gaussian mutual information is significantly smaller (resp. slightly smaller or similar) and has a significantly lower (resp. slightly lower or similar) variance than mutual information for durations above (resp. below) 100 ms. These variances indicate that the nonstationary Gaussian source model is more appropriate for audio sources than the stationary non-Gaussian model.

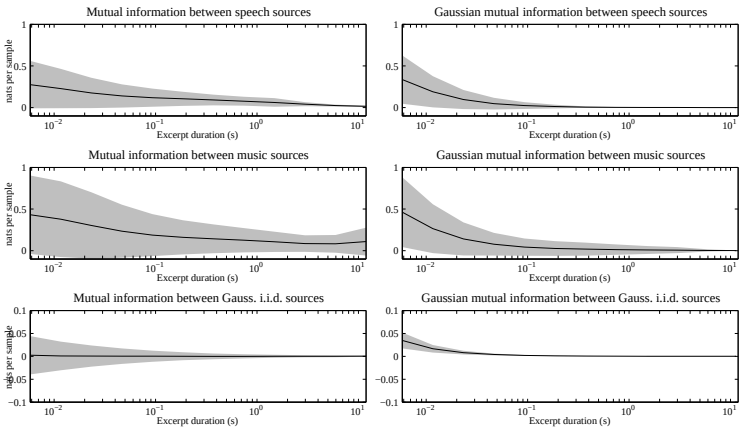

Fig. 2. Mutual information and Gaussian mutual information between the source signals in [7] and between Gaussian i.i.d. signals vs excerpt length. The plain lines and the gray areas resp. denote the mean and the spread (one standard deviation) of the measured values.

In order to show the influence of the dependency measures on the performance of classical ICA methods, we consider the above excerpts, we mix them with the identity matrix and we run the parallel version of FastICA [8][4] and the Pham-Cardoso algorithm [6], since these methods are resp. based on non-Gaussian stationary and Gaussian nonstationary dependency measures. The performance index is the well-known Signal-to-Interference Ratio (SIR) [9]. Figure 3 shows the performance of the above BSS methods, with respect to the excerpt length. Note that in some cases, we found ill-conditioning problems with the Pham-Cardoso algorithm: in the joint-diagonalization procedure of this BSS

[4] FastICA matlab code is accessible at http://www.cis.hut.fi/projects/ica/fastica/.

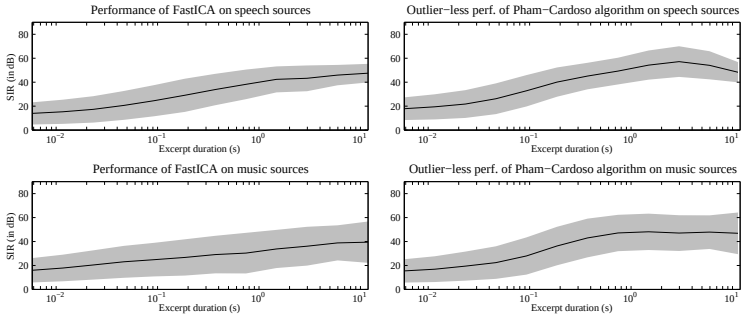

Fig. 3. Performance of ICA methods on the source signals in [7], mixed by the identity matrix. The plain lines correspond to mean SIR (in dB) while the gray areas show the spread.

method, the inverse mixing matrix is estimated in an iterative procedure. We found cases when the determinant of $\widehat{\mathbf{R}}_s(q)$ in (2) is equal to zero, thus making the iterative procedure stop early and the final estimated separation matrices equal to its initial value. Since it is initialized as the identity matrix (which is also our mixing matrix), the corresponding SIR is very large, yielding an aberrant value. These outliers are removed in Fig. 3. The performance of FastICA and of the outlier-less version of the Pham-Cardoso algorithm is in agreement with the above dependency measures, except in the case of speech sources separated by the Pham-Cardoso algorithm, since the performance slightly decreases for the longest durations. Indeed, when the size of the excerpts decreases, the dependency between the sources increases and the SIR obtained by the BSS methods significantly decrease. Since the mean dependency measures between sources are high for short-duration excerpts, one could expect the corresponding SIR to be low. Figure 3 shows that it is close to 20 dB, which still yields significant source enhancement. Lastly, except for the shortest music source excerpts, the Pham-Cardoso algorithm yields the highest mean SIR, which confirms the above comments on the dependency measures.

As a consequence of the above analysis, for the sake of brevity, we only study the Gaussian mutual information measure and the corresponding Pham-Cardoso BSS method below, since they are resp. more appropriate for speech and music source signals than the non-Gaussian dependency estimators and the ICA approaches based on these measures.

3 Independence and ICA Performance in Time-Frequency Domain

As explained in Sect. 1, the performance of frequency-domain convolutive BSS has been discussed in [5] by Araki *et al.* with respect to the time-frequency win-

dow length of the time-frequency transform and the length of the mixing filter impulse responses. Indeed, they show that the performance of frequency approaches highly depends on the temporal width of the time-frequency windows. Here, we extend to the time-frequency domain the procedure proposed in Sect. 2. This work differs from [5] since Araki *et al.* only measured correlation between sources (instead of the dependency measures used in their BSS method) and did not consider the nature of source signals. However, they took into account the length of the impulse responses of the mixing filters, which is not studied here, for the sake of brevity. As stated above, we only consider the Gaussian mutual information, since it is more appropriate for audio sources than the mutual information. We still use the signals [7] tested in the previous section and we compute their short-time Fourier transform (STFT), defined as

$$S_i(t,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} s_i(t') h(t'-t) e^{-j\omega t'} dt' \qquad i = 1,2, \tag{3}$$

where $h(t'-t)$ is a windowing function centered on time t. In these tests, the length of this windowing function geometrically increases from 2^7 to 2^{13} samples. For each STFT window length, for each pair of source signals and for each frequency bin, we compute the Gaussian mutual information. In Fig. 4, we show the mean values and the spread of the considered dependency measure, over the frequency bins, with respect to the length of the windowing function used in (3). Like in Sect. 2, we carry out the same experiment for thirty pairs of independent Gaussian white noises. Contrary to the previous analysis in time domain, here, audio and noise mean dependency measures are of the same order of magnitude. Moreover, the estimated dependency of the sources increases when the size of $h(.)$ increases. This phenomenon, explained in [5], can be summarized as follows: when the STFT window size is high, the number of time-frequency samples in each frequency bin is small and estimating correctly the statistics becomes harder. However, even if all the mean dependency measure values decrease with the STFT size, the ratio between music and speech mean dependency measures (not presented here) significantly increases when the STFT size decreases. This means that music sources present more dependencies than speech for low STFT sizes. Lastly, music and speech sources provide a larger variance than white noise. In particular, the largest variance is obtained with music source signals, which is coherent with the results obtained in Sect. 2.

We then apply the Pham-Cardoso algorithm on each frequency bin of the time-frequency transforms of the sources. In order to analyze the sensitivity of this algorithm to dependency only, we do not generate "real" convolutive mixtures. Indeed, the approximation of convolution by complex-valued multiplication in each frequency bin also affects performance. We avoid measuring this effect by actually generating mixtures from a complex-valued mixing matrix in each bin. Since the considered ICA algorithm is equivariant, its performance does not depend on the value of the chosen matrix. Therefore we simply choose the identity matrix, as in Sect. 2. In order to handle the band-to-band permutation effects, we compute the SIR on the output signals obtained for each

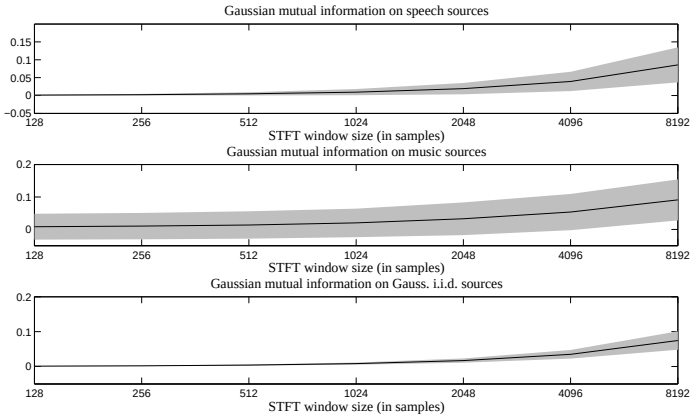

Fig. 4. Gaussian mutual information between the time-frequency transforms of the source signals in [7]. The plain lines correspond to mean value while the gray areas show the spread.

frequency bin. Figure 5 shows the mean SIR, over all the frequency bins and the sets of sources, with respect to the STFT window length. The results are in agreement with those in Fig. 4: both classes of signals yield decreasing SIR when the STFT window size increases, which is again in agreement with [5]. For the shortest STFT sizes, the performance obtained with speech sources is higher than for music ones, which is in agreement with the above analysis on dependency measures. Moreover, the variance of the SIR obtained with speech sources is somewhat lower than the one obtained with music sources. Lastly, note that the mean SIR are much higher than those obtained by frequency-domain convolutive ICA algorithms in the literature (which are generally around 10 dB). This is due to the band-to-band permutation problem and to the length of the impulse response of the mixing filters, which have been occulted here.

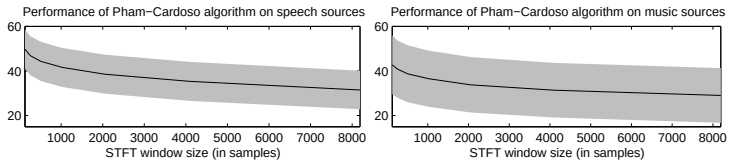

Fig. 5. Performance of the Pham-Cardoso method for the source signals in [7], mixed by the identity matrix in the framework of frequency domain convolutive BSS. The plain lines correspond to the mean SIR (in dB) while the gray areas show the spread.

4 Conclusion

In this paper, we studied the validity of the independence assumption for speech or music source mixtures, with respect to the excerpt size (resp. the STFT window size) in the time (resp. time-frequency) domain. Starting from previous work stating that music sources are dependent [2] for short durations, we looked for statistical differences between speech and music dependency measures in several configurations. We finally showed that these classes of sources almost yield the same mean behaviour for long excerpt durations and high STFT window sizes. However, for both linear instantaneous and convolutive mixture models, the variance of the dependency measures is significantly higher for music than for speech. Moreover, in the time domain, we showed that these classes of sources are separable by ICA for long-enough durations while, for short-time excerpts, the mean dependencies are high, which implies that ICA methods are not appropriate in these cases, which is in agreement with [2,3]. This limitation may be solved thanks to BSS methods not based on independence, e.g. [2,10]. Moreover, even if the independence assumption is met, sparse algorithms often separate much better speech and music sources than classical ICA methods (see e.g. [11]). It could be interesting to compare the performance of sparse approaches with respect to the degree of sparsity of the source signals, for several classes of sources.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley-Interscience, New York. (2001)
2. Abrard, F., Deville, Y.: Blind separation of dependent sources using the "TIme-Frequency Ratio Of Mixtures" approach. In: Proc. Int. Symp. on Signal Processing and its Applications (ISSPA). (2003) 81–84
3. Smith, D., Lukasiak, J., Burnett, I.S.: An analysis of the limitations of blind signal separation application with speech. Signal Processing **86**(2) (2006) 353–359
4. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information Physical Review E **69**(6) (2004) preprint 066138
5. Araki, S., Makino, S., Mukai, R., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. IEEE Trans. on Speech and Audio Processing **11**(2) (2003) 109–116
6. Pham, D.T., Cardoso, J.-F.: Blind separation of instantaneous mixtures of nonstationary sources. IEEE Trans. on Signal Processing **49**(9) (2001) 1837–1848
7. Vincent, E., Gribonval, R., Plumbley, M.D.: Oracle estimators for the benchmarking of source separation algorithms. Signal Processing **87**(8) (2007) 1933–1950
8. Hyvärinen, A.: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Trans. on Neural Networks **10**(3) (1999) 626–634
9. Mansour, A., Kawamoto, M., Ohnishi, N. A survey of the performance indexes of ICA algorithms. Proc. of Int. Conf. on Modelling, Identification, and Control (2002)
10. Caiafa, C.F., Proto, A.N.: Separation of statistically dependent sources using an L^2-distance non-Gaussianity measure. Signal Processing **86** (2006) 3404–3420
11. Deville, Y., Puigt, M.: Temporal and time-frequency correlation-based blind source separation methods. Part I: Determined and underdetermined linear instantaneous mixtures. Signal Processing **87**(3) (2007) 374–407