Extension of the ”TIme-Frequency Ratio Of Mixtures” blind source separation method to more than 2 channels

Frédéric Abrard, Yannick Deville
Laboratoire d'Acoustique, de Métrologie et d'Instrumentation
Université Paul Sabatier, Bât. 3R1B2, 118 route de Narbonne, 31062 Toulouse Cedex, FRANCE
abrard@i2e.fr - ydeville@cict.fr

Abstract

*In a recent paper, we proposed a new blind source separation (BSS) method, which uses time-frequency (*TF*) information to extract two source signals from two linear instantaneous mixtures of these sources. In this new paper, we introduce an extension of the latter method, intended for the general situation when N mixtures of N source signals are available. Unlike previously reported* TF *BSS methods, the proposed approach only requires slight differences in the* TF *distributions of the considered signals: it mainly requests the sources to be ”visible”, i.e. to each occur alone in one local area of the* TF *plane. By using* TF *ratios of mixed signals, it automatically determines these single-source* TF *areas and identifies the corresponding parts of the mixing matrix. We present in detail the proposed method and give experimental results concerning mixtures of speech and music signals, thus showing that this approach yields very good performance.*

I. INTRODUCTION

Blind source separation (BSS) consists in estimating a set of N unknown sources from N observations resulting from the mixture of these sources through unknown propagation channels. Denoting the mixing operator by $\mathcal{A}$, the relationship between the sources and observations reads $\underline{x} = \mathcal{A}\underline{s}$, where the vector $\underline{s} = [s_1,\ s_2,\ \ldots,\ s_N]^T$ contains the unknown sources while $\underline{x} = [x_1,\ x_2,\ \ldots,\ x_N]^T$ represents the observations. We here only consider linear instantaneous mixtures, so that the operator $\mathcal{A}$ corresponds to a scalar matrix.

Traditional Independent Component Analysis (ICA) approaches basically aim at separating the sources by combining the observations so that the output signals are independent [1] which means that the fundamental assumption of ICA techniques is that the sources must be independent. Moreover, most of these approaches can only separate stationary non-Gaussian signals. Because of these limitations, poor performance is often obtained when dealing with real sources, like audio signals, which do not match those requirements. Some authors [2]-[8] have proposed different approaches which take advantage of the non-stationarity of such sources in order to achieve better performance than classical methods for this type of signals. However, the approaches presented in [2]-[4] do not apply to the underdetermined case, as they then yield signals which are still mixtures of all source signals. To overcome the latter restriction, we proposed an original concept for the underdetermined case [5]-[8]. This method is efficient but requires the sources to have specific stationarity properties. Audio signals, for example, are not well suited to this approach and another solution is required for them. The method that we introduce in this paper solves this problem.

A few authors [9],[10] proposed BSS methods which use *time-frequency* (*TF*) information. However, their approaches are quite complex and require high computational load. Recently, a *TF* method for time-delayed mixtures has been presented [11] and also tested with convolutive mixtures, using real-time computation [12]. But this method ideally requires the sources to be disjoint orthogonal in the *TF* plane, i.e. only one source should occur in each *TF* window, which is quite restrictive.

In a recent paper [13] we proposed a new BSS method which uses (*TF*) information to extract two source signals from two linear instantaneous mixtures of these sources. In this new paper, we introduce an extension of the latter method, intended for the general situation when N mixtures of N source signals are available. In this method, we exploit the *TF* information derived from the observations in a different way than in the approaches reported in the literature, in order to automatically determine

some *TF* windows where a single source occurs. Unlike in [11]-[12], we need the sources to occur alone in only a small area of the *TF* plane and we do not perform the source reconstruction from some specific parts of the *TF* plane, nor do we need any iterative algorithm.

This paper is organized as follows. In Section II, we present the *TF* method that we recently proposed for the simple configuration involving two mixtures of two sources, in order to introduce the concepts and notations which are required in the remainder of this paper. We then extend this method to the case of N sources and N observations in Section III. Experimental results on mixtures of audio signals are presented in section IV. We eventually draw various conclusions from this investigation in Section V.

II. Basic case: Two mixtures of two sources

A. Problem statement

We here consider the following linear instantaneous mixture[1] of two real-valued sources:

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \end{cases} \tag{1}$$

where the coefficients a_{ij} of the mixing matrix A are real, constant and different from zero.
The separation of the sources s_i can classically only be performed up to a scale factor and a permutation [1] and BSS may thus be seen as a method for finding an estimate of $\tilde{A}^{-1} = \Lambda P A^{-1}$, where Λ and P are resp. arbitrary diagonal and permutation matrices. Inside this class of matrices, we here focus on:

$$\tilde{A}^{-1} = \begin{bmatrix} 1 & 1 \\ 1/c_{12} & 1/c_{22} \end{bmatrix}^{-1} \tag{2}$$

where

$$c_{12} = \tfrac{a_{11}}{a_{21}}, \quad c_{22} = \tfrac{a_{12}}{a_{22}}, \tag{3}$$

correspond to the *cancelling coefficient values* introduced and used in [13] to cancel one source from the observations by using a specific cancelling structure. To go further, we now use these cancelling coefficient values to build the inverse matrix (2) which, applied to $\underline{x}(t)$, yields the output vector:

$$\underline{y}(t) = \tilde{A}^{-1}\underline{x}(t) = [a_{11}s_1(t), a_{12}s_2(t)]^T. \tag{4}$$

Applied to $\underline{x}(t)$ the matrix (2) thus cancels the source s_2 and s_1 resp. in the first and second output.

We proposed in [13] a method to find these "cancelling coefficient values" based on time-frequency analysis that we recall hereafter.

B. Time-frequency analysis

1) Definition of the time-frequency tool: Inside the large set of *TF* tools developed outside the scope of BSS [14], [15] we here restrict ourselves to the simplest one, i.e. the *short-time Fourier transform (STFT)*. We choose this transform because it does not have any interference terms, which is crucial for our approach, and is efficiently computed thanks to FFT algorithms.

Considering each mixed signal $x_i(\tau)$, the STFT of x_i is given by [16]:

$$X_i(t,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x_i(\tau) h(\tau - t) e^{-j\omega\tau} d\tau. \tag{5}$$

where $h(\tau - t)$ is a shifted real-valued window function, centered at time t. $X_i(t,\omega)$ is the contribution of signal x_i in the short time and frequency windows resp. centered on t and ω.

It should be noted that the STFT is initially defined for deterministic signals and is indeed applied in such a framework in this paper: even if the considered sources and observations are random processes, the STFTs used hereafter only concern a single, and therefore deterministic, realization of these signals (which is requested to satisfy the assumptions defined below).

[1]The mixtures are assumed to be non-degenerate throughout this paper.

2) Exploiting time-frequency information: We now show how *TF* analysis may be used to identify the cancelling coefficient values (3). We then introduce an automatic method for finding the appropriate *TF* areas. To this end, we request the following assumptions:

Assumption 1: The mixing matrix A is such that $a_{ij} \neq 0,\ \forall\, i,\ j$ and the power of each source is non negligible at least at some times t.

Assumption 2: For each source s_i, there exist some adjacent *TF* windows (t_j, ω_k) where only s_i occurs, i.e. where[2]: $S_l(t_j, \omega_k) \ll S_i(t_j, \omega_k),\ \forall\, l \neq i$.

Our BSS method is then based on the complex ratio:

$$\alpha(t_j, \omega_k) = \frac{X_1(t_j, \omega_k)}{X_2(t_j, \omega_k)}, \tag{6}$$

which is computed for each *TF* window. Taking into account Equ. (1) and (5) leads to:

$$\alpha(t_j, \omega_k) = \frac{a_{11} S_1(t_j, \omega_k) + a_{12} S_2(t_j, \omega_k)}{a_{21} S_1(t_j, \omega_k) + a_{22} S_2(t_j, \omega_k)}. \tag{7}$$

Therefore, if one source does not have any component in the *TF* window (t_j, ω_k), then $\alpha(t_j, \omega_k)$ is equal to the cancelling coefficient value, among c_{12} and c_{22} defined in (3), which makes it possible to extract this source. This situation when sources only disappear in some areas of the *TF* plane is very frequent. The problem is now to find a method to determine such areas. The following assumption is required to this end:

Assumption 3: When several sources occur in a given set of adjacent *TF* windows, they should vary so that $\alpha(t, \omega)$ does not take the same value in all these windows.

Thus, if only source $s_i(t)$ is present in several time-adjacent windows[3] (t_j, ω_k) then $\alpha(t_j, \omega_k)$ is constant and equal to c_{i2} over these successive windows, whereas it takes different values over these windows if both sources are present and if Assumption 3 is met.

To exploit this phenomenon, we compute the sample variance of the complex ratio $\alpha(t, \omega)$ on series Γ_q of M short half-overlapping time windows corresponding to adjacent t_j, applying this approach to each frequency ω_k. We resp. define the sample mean and variance of $\alpha(t, \omega)$ on Γ_q and ω_k by $\overline{\alpha}(\Gamma_q, \omega_k) = \frac{1}{M} \sum_{j=1}^{M} \alpha(t_j, \omega_k)$ and $var[\alpha](\Gamma_q, \omega_k) = \frac{1}{M} \sum_{j=1}^{M} |\alpha(t_j, \omega_k) - \overline{\alpha}(\Gamma_q, \omega_k)|^2$.

If e.g. $S_2(t_j, \omega_k) = 0$ for these M windows, then (7) shows that $\alpha(t_j, \omega_k)$ is constant over them, so that its variance $var[\alpha](\Gamma_q, \omega_k)$ is equal to zero. Conversely, under Assumption 3, if both $S_1(t_j, \omega_k)$ and $S_2(t_j, \omega_k)$ are different from zero then $var[\alpha](\Gamma_q, \omega_k)$ is significantly different from zero.
So, by searching for the lowest value of $var[\alpha](\Gamma_q, \omega_k)$ vs all the available series of windows (Γ_q, ω_k), we directly find a *TF* domain (Γ_q, ω_k) with only one source. The corresponding value c_{i2} which cancels this source is then estimated by the mean $\overline{\alpha}(\Gamma_q, \omega_k)$. We find the second cancelling coefficient value c_{i2} by searching for the next lowest value of $var[\alpha](\Gamma_q, \omega_k)$ vs (Γ_q, ω_k) associated to a significantly different value of $\overline{\alpha}(\Gamma_q, \omega_k)$ using a threshold set to the minimum difference that we request between the two values in (3). We thus obtain estimates of the two cancelling coefficient values defined in (3). The separated signals are then derived from these values by using i) either the original version of the approach based on individual source extractions that we proposed in [13] or ii) its new version based on the matrix (2). If the lowest value of the ratio variance is obtained when s_2 is zero this yields (3) and (4). Otherwise a permutation occurs in (3) and (4).

III. Extension to N mixtures of N sources

We now show how the above method may be extended to the case when N mixtures of N source signals are available. For the sake of clarity, we first consider an intermediate situation.

[2]This situation is e.g. common for speech or music signals: the formants of speakers or instruments are located in *TF* areas which do not overlap completely.

[3]The same concept may be applied to frequency-adjacent windows.

A. $N > 2$ sources, 2 observations

As an intermediate step, let us consider the situation when 2 observed mixtures are available, but they now contain more than 2 source signals. The observations then become:

$$\begin{cases} x_1(t) = \sum_{m=1}^{N} a_{1m} s_m(t) \\ x_2(t) = \sum_{n=1}^{N} a_{2n} s_n(t) \end{cases} \tag{8}$$

It is easily shown that applying to the vector $\underline{x}(t)$ any 2×2 "partial inverse" matrix

$$\tilde{A}^{-1} = \begin{bmatrix} 1 & 1 \\ 1/c_{i2} & 1/c_{j2} \end{bmatrix}^{-1} \tag{9}$$

where $c_{i2} = \frac{a_{1i}}{a_{2i}}$ provides two different outputs with resp. cancellation of s_j and s_i.

The BSS method defined in Subsection II-B is therefore straightforwardly extended to the current case, but then leads to a *partial* separation, i.e. to the cancellation of only one of the existing sources in each output signal. This is of high practical interest in signal enhancement applications anyway, as this method gives an efficient solution for removing the contribution of an undesirable source.

B. General case: N sources, N observations

1) Coherence of the time-frequency maps:

Due to Assumption 1, the areas (Γ_q, ω_k) where a given source appears alone in observations are the same for all observations. We call this phenomenon the "*coherence of the* TF *maps*". Thanks to this *coherence*, single-source areas may be detected for most mixing matrices by analyzing the variance of the ratio $\alpha(t,\omega) = X_i(t,\omega)/X_j(t,\omega)$ associated to only one arbitrary pair of observations: here again, this variance is low in and only in single-source areas under Assumption 3.

An exception to this principle appears when the number of observations is higher than 2 however: in areas (Γ_q, ω_k) where several sources are active, $\alpha(t,\omega)$ may have a low variance for some pairs of observations, because the corresponding subset of mixing coefficients results in proportional observations in these areas. For a given area, this phenonemon may not occur for all pairs of observations however, otherwise the mixing matrix would be degenerate. This case, which only concerns very specific mixing matrices, is therefore handled by performing variance analyses for all pairs of observations $(x_1(t), x_j(t))$. We skip this specific case hereafter and therefore only consider a single variance analysis, thus introducing a fast BSS method.

2) Fast resolution for N mixtures of N sources:

We here suppose that N observed mixtures of N sources are available and that the above assumptions are still met. As explained in Subsection III-B.1, we first perform a single variance analysis with two observations. This yields all the *TF* areas where only one source occurs for all the observations. We then adapt the approach of Subsection III-A to each pair of observations $(x_1(t), x_j(t))$. We thus compute the mean of the ratio $X_1(t,\omega)/X_j(t,\omega)$ in the area given by the variance analysis where only s_i exists, which yields the value $c_{ij} = a_{1i}/a_{ji}$. Using these values, we then build a matrix $\tilde{A}^{-1}$ which achieves global inversion up to a scale factor, i.e:

$$\tilde{A}^{-1} = \begin{bmatrix} 1 & \cdots & 1 \\ 1/c_{12} & \cdots & 1/c_{N2} \\ \vdots & & \vdots \\ 1/c_{1N} & \cdots & 1/c_{NN} \end{bmatrix}^{-1}, \tag{10}$$

which yields: $\underline{y}(t) = \tilde{A}^{-1}\underline{x}(t) = [a_{11}s_1(t), \ldots, a_{1N}s_N(t)]^T$. This efficient method leads to a complete BSS in one step.

IV. Experimental results

To illustrate the ability of this method to handle the general case of square mixtures we considered 4 sources and 4 observation. To make things harder, 3 sources are different sentences recorded from the same speaker, which means that harmonic components are nearly in the same *TF* locations. The 4^{th}

source is recorded from a singer with components spread in the whole *TF* plane. We used a sampling frequency of 22050 Hz. The chosen mixing matrix has a small determinant of 0.06 and Table I gives the input SNR's. Fig. 1 to 4 show the 256-sample spectrograms of each source. As an example, we analyzed the variance of $\alpha(n_j, \omega_k)$ for $M = 8$ on 0.45 s of signals (10000 samples), which took approximately 3 s with matlab code on a 1GHz PIII, and plotted in Fig. 5 the result $\frac{1}{var[\alpha](\Gamma_q, \omega_k)}$. One can easily see that there exist some areas with low variance, appearing as peaks and corresponding to windows where only one source occurs. Table II gives the output SNR's obtained for each source on the different outputs. We see that each source has been successfully extracted with an average SNR of 38 dB. We show in Table III results with different sizes of STFT windows and number M of these windows for the variance analysis. As we can see, separation is always achieved with good SNR's, even for short windows.

V. Discussion and conclusion

In this paper, we proposed a simple and efficient method for solving the linear instantaneous BSS problem with N sources and N observations.

This approach is based on the TIme-Frequency version of Ratios Of Mixtures of source signals, and is therefore called "TIFROM". It mainly relies on the assumption that the sources are "visible", i.e. that each of them occurs alone (as opposed to the other sources) in at least one local area in the *TF* plane. It automatically determines such an area and then derives coefficients which allow one to cancel the contributions of this source from the observed signals.

This approach yields major advantages over classical methods. Especially, it applies to stationary and non-stationary signals. It is also more attractive than *TF* BSS methods previously reported by other authors, because it sets much less restrictive constraints on the *TF* distributions of the sources, it is based on simple principles yielding low computational load and it performs BSS without any convergence issues.

Some restrictions may appear when applying this method to a large number of sources, as their chance to be visible then decreases. However, experimental tests performed with mixtures of 4 sources having similar *TF* distributions show that we can find appropriate *TF* areas for all of them and then obtain good values of output SNR's.

Our future investigations will esp. concern this case involving many sources and the extension of the proposed approach to convolutive mixtures.

References

[1] J. F. Cardoso, "Blind signal separation: statistical principles," in *Proc. of the IEEE*, vol. 86, no. 10, Oct. 1998, pp. 2009–2025.

[2] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of non-stationary sources," *IEEE Transaction on Signal Processing*, October 2000.

[3] A. Hyvarinen, "Blind source separation by nonstationarity of variance: a cumulant-based approach," *IEEE Trans. on Neural Networks*, vol. 12, no. 6, pp. 1471–1474, November 2001.

[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transaction on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[5] Y. Deville and M. Benali, "Differential source separation: concept and application to a criterion based on differential normalized kurtosis," in *Proceedings of EUSIPCO*, Tampere, Finland, September, 4-8, 2000.

[6] Y. Deville, F. Abrard, and M. Benali, "A new source separation concept and its validation on a preliminary speech enhancement configuration," in *Proceedings of CFA2000*, Lausanne, Switzerland, September 3-6, 2000, pp. 610–613.

[7] F. Abrard, Y. Deville, and M. Benali, "Numerical and analytical solution to the differential source separation problem," in *Proceedings of EUSIPCO*, Tampere, Finland, September, 4-8, 2000.

[8] Y. Deville and S. Savoldelli, "A second-order differential approach for underdetermined convolutive source separation," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), session MULT-P2*, Salt Lake City, USA, May 7-11 2001.

[9] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, November 1998.

[10] L. Giulieri, N. Thirion-Moreau, and P. Y. Arquès, "Blind source separation using bilinear and quadratic time-frequency representations," in *Proceedings of ICA 2001*, San Diego, December 9-13, 2001.

[11] A. Jourjine, S. Rickard, and . Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *Proceedings of ICASSP 2000*, vol. 6, Istanbul, Turkey, June, 6-9, 2000, pp. 2986–2988.

[12] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proceedings of ICA 2001*, San Diego, CA, December, 9-13, 2001.

[13] F. Abrard, Y. Deville, and P. R. White, "A new source separation approach for instantaneous mixtures based on time-frequency analysis," in *Proceedings of ECM^{2}S*, Toulouse, France, May 2001.

[14] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Magazine*, vol. 9, pp. 21–67, April 1992.
[15] L. Cohen, "Time-frequency distributions - a review," in *Proceedings of the IEEE*, vol. 77, No. 7, July 1989, pp. 941–979.
[16] ——, *Time-frequency analysis*. Englewood Cliffs, New Jersey: Prentice hall PTR, 1995.

	x_1	x_2	x_3	x_4
s_1	0.4	-9.4	-3.4	-12
s_2	-5.1	-2.5	-3.6	-9.6
s_3	-5.9	-7.5	0.61	-9.1
s_4	-2.5	-4.8	-4.6	-8.1

TABLE I
INPUT SNR'S (DB)

	out 1	out 2	out 3	out 4
s_1	-62	-48	39	-70
s_2	-36	-38	-45	43
s_3	-39	36	-40	-50
s_4	34	-43	-67	-44

TABLE II
OUTPUT SNR'S (DB)

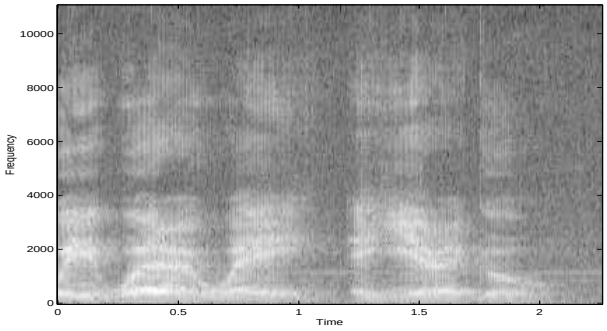


Fig. 1. Spectrogram of s_1 for 256-sample STFT windows.

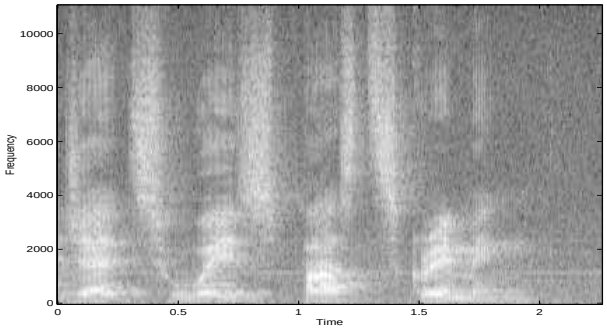


Fig. 2. Spectrogram of s_2 for 256-sample STFT windows.

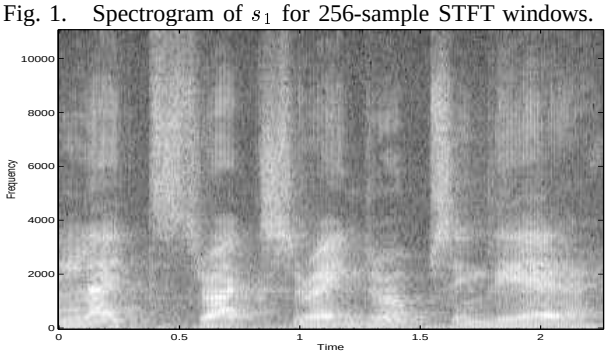


Fig. 3. Spectrogram of s_3 for 256-sample STFT windows.

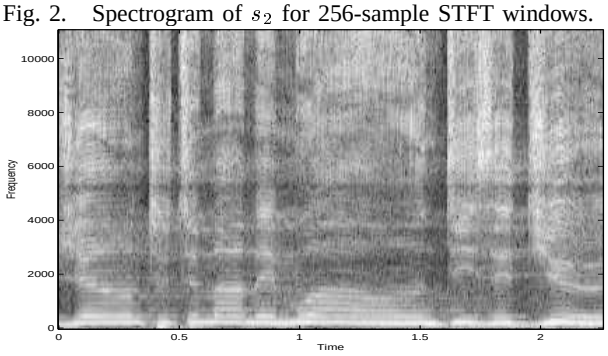
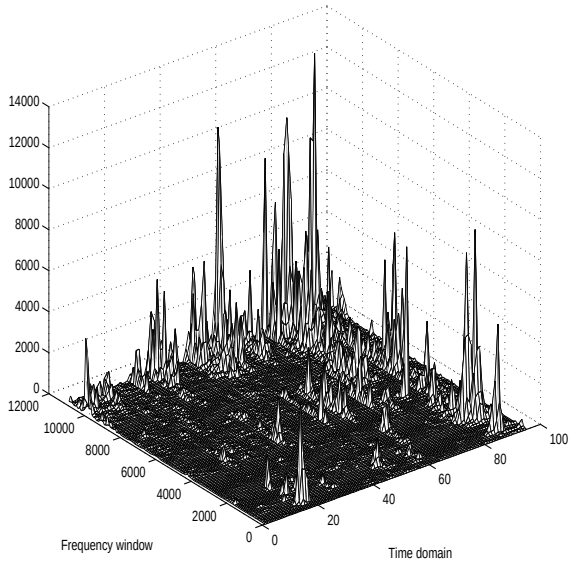


Fig. 4. Spectrogram of s_4 for 256-sample STFT windows.



		N_{STFT}		
M		64	128	256
4	s_1	48	44	41
	s_2	32	36	46
	s_3	32	35	45
	s_4	24	45	46
8	s_1	45	48	39
	s_2	34	41	43
	s_3	39	43	36
	s_4	32	34	34
12	s_1	54	44	42
	s_2	43	46	48
	s_3	40	49	50
	s_4	32	45	35

TABLE III

Fig. 5. Time-Frequency representation of $\frac{1}{var[\alpha](\Gamma_q,\omega_k)}$. Axes OUTPUT SNR (DB) VS N_{STFT} AND M FOR EACH SOURCE. units : Time window indices, corresponding to [0 s, 0.45 s]. Frequency window indices, corresponding to [0 Hz, 22.05 kHz].