FROM BLIND SOURCE SEPARATION TO BLIND SOURCE CANCELLATION IN THE UNDERDETERMINED CASE: A NEW APPROACH BASED ON TIME-FREQUENCY ANALYSIS

Frédéric Abrard and *Yannick Deville*

Laboratoire d'Acoustique,
de Métrologie et d'Instrumentation
Université Paul Sabatier
118 route de Narbonne
31062 Toulouse cedex, France
abrard@cict.fr, ydeville@cict.fr

Paul White

Signal Processing and Control Group
Institute of Sound and Vibration Research
University of Southampton
Highfield SO17 1BJ
England
prw@isvr.soton.ac.uk

ABSTRACT

Many source separation methods are restricted to non-Gaussian, stationary and independent sources. This yields some problems in real applications where the sources often do not match these hypotheses. Moreover, in some cases we are dealing with more sources than available observations which is critical for most classical source separation approaches.
In this paper, we propose a new simple source separation method which uses time-frequency information to cancel one source signal from two observations in linear instantaneous mixtures. This efficient method is directly designed for non-stationary sources and applies to various dependent or Gaussian signals which have different time-frequency representations. Its other attractive feature is that it performs source cancellation when the two considered mixtures contain more than two sources.
Detailed results concerning mixtures of speech and music signals are presented in this paper.

1. INTRODUCTION

We first consider the following mixture model:

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

where the coefficients a_{ij} are real and constant.
Our goal is to find a method for separating two source signals s_1 and s_2 from the two observations x_1 and x_2 without knowing the mixing coefficients a_{ij} nor the sources x_i.
This problem is called *Blind Source Separation (BSS)* and is well known in the signal processing community.
Writing Equ. (1) in matrix notation $X = AS$, this problem is equivalent to finding an inverse matrix B such that $B = \lambda PA^{-1}$ where P is a permutation matrix and λ is a diagonal matrix [1].
One can find a review of many methods for achieving this separation in [1]. Most of them are *statistic-based* methods including an adaptive part and can only be applied to specific signals like stationary and non-Gaussian signals. Moreover, these methods need the source signals to be independent and often fail when more sources than sensors are present in the observations. Especially, we recently proposed an approach based on 4^{th}-order normalized cumulants (i.e. kurtosis) [2] allowing one to solve the problem when the number of sources is equal to the number of observations. This method consists in finding a linear combination of the two observations

$$y(t) = x_1(t) - c.x_2(t) \quad (2)$$

which achieves the extraction of one source up to a scale factor. The proper *separating coefficients* c_i for extracting $s_1(t)$ or $s_2(t)$ by means of Equ. (2) are respectively:

$$c_1 = \frac{a_{12}}{a_{22}} \quad c_2 = \frac{a_{11}}{a_{21}} \quad (3)$$

We also proposed a related solution for the underdetermined case by cancelling the influence of the stationary sources during the adaptation step in order to achieve a partial source separation [3], [4], [5]. These methods are efficient but the sources must be non-Gaussian, independent with a special stationarity.
We show in this paper that these restrictions can be reduced if we use the time and frequency information of the signals. A few authors [6], [7] proposed solutions using time-frequency information but their approaches are complex and require high-computational load. With the same separation structure as in Equ. (2), we propose here a new simple time-frequency method for cancelling one source with less restrictions than with classical methods.

2. PRELIMINARY IDEA: TEMPORAL ANALYSIS

If we can find sections in the time domain where $x_1(t)$ and $x_2(t)$ contain only the contribution of one source, we can easily find the separating coefficient values c_i that we introduced in Equ. (3). For example if we can find a time t_n such that $s_2(t_n) = 0$, then (1) yields:

$$\begin{cases} x_1(t_n) = a_{11}s_1(t_n) \\ x_2(t_n) = a_{21}s_1(t_n) \end{cases} \quad (4)$$

By computing the ratio

$$\frac{x_1(t_n)}{x_2(t_n)} = \frac{a_{11}}{a_{21}} \quad (5)$$

we directly obtain the value c_2 of c which extracts the source $s_2(t)$. This means that we theoretically only need a source to disappear at time t_n to find a separating coefficient.

This is a really simple source separation method but, unfortunately, it is usually hard to find an instant or time interval where only one source occurs. To overcome this problem we propose a new approach exploiting the time-frequency domain.

3. TIME-FREQUENCY ANALYSIS

In the previous section, we presented a technique for finding the separating coefficient if one source "disappears" over a known short time interval. But we need to find a more general method allowing one to solve this problem if both sources are simultaneously present or if one does not know when these sources disappear.
To this end, we use and request the following assumptions:

1. The time-frequency transform of each source must be different for time-adjacent time-frequency windows[1].

2. There must exist some time-frequency windows where only one source is present[2].

Many powerful time-frequency methods have been developed during the last fifty years with different application fields. One can find most of them with detailed references in [8], [9], [10], [11].
To avoid the interference areas present in the 2^{nd} and higher order existing methods, the most relevant starting point to solve our problem is to use the simple short time Fourier transform of the observations as defined in [10]. We first multiply each mixed signal $x_i(\tau)$ by a shifted Hanning window function $h(\tau - t)$, centered at time t, to produce the modified signal:

$$x_i(t, \tau) = x_i(\tau)h(\tau - t) \quad (6)$$

This new function is now a function of two times, the fixed time we are interested in, t, and the running time τ.
We then compute the short time Fourier transform of each $x_i(t, \tau)$, i.e:

$$X_i(t, \omega) = \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau} x_i(\tau)h(\tau - t)d\tau \quad (7)$$

Our goal is now to find some time-frequency domains where only one source occurs. To this end we introduce the complex ratio:

$$\alpha(t, \omega) = \frac{X_1(t, \omega)}{X_2(t, \omega)} \quad (8)$$

This ratio is computed for each time and angular frequency window. With Equ. (1), this leads to:

$$\alpha(t, \omega) = \frac{a_{11}S_1(t, \omega) + a_{12}S_2(t, \omega)}{a_{21}S_1(t, \omega) + a_{22}S_2(t, \omega)} \quad (9)$$

One can easily see that if one source does not have any component at (t, ω), i.e on the Hanning time window and frequency window respectively centered on t and ω, then $\alpha(t, \omega)$ is real and equal to the value of the separating coefficient c for extracting this source. For example if $S_2(t, \omega)$ is missing then $\alpha(t, \omega)$ becomes:

$$\alpha(t, \omega) = \frac{a_{11}}{a_{21}} \quad (10)$$

which is the correct coefficient to extract $s_2(t)$ with Equ. (2). This situation, when sources have slightly different time-frequency representations is more frequent than the case when one source disappears during a time period. For example the time-frequency properties of two people speaking at the same time are different.
We denote $(\Theta, \Omega) = \bigcup_n \{(t_n, \omega_n) \ / \$ only one source occurs at (t_n, ω_n) $\}$.

Now the remaining question is how can we find these (t_n, ω_n) domains ?
Our idea is that each value $\alpha(t_n, \omega_n)$ is ideally equal to c_1 or c_2 as $(t_n, \omega_n) \in (\Theta, \Omega)$, whereas it takes different values in all the other regions $(t, \omega) \notin (\Theta, \Omega)$. Especially, if only source $s_1(t)$ is present in several successive (t_n, ω_n) then $\alpha(t_n, \omega_n)$ is constant and equal to c_2 over these successive windows, whereas it successively takes different values if both sources are present *AND* if their time-frequency representations are not constant. To exploit this, we compute the statistical variance of $\alpha(t, \omega)$ on a limited series Γ_k of M short half-overlapping time windows corresponding to

[1] Due to statistical fluctuations, even white noise signals with theoretically constant power spectrum densities satisfy this assumption for short time windows in practice.

[2] This situation is really common in speech or music for example. The formants of a same or different speaker/instrument are located in different time-frequency areas depending on the produced sound.

t_i, and this for each frequency window ω_j. We resp. define the mean and variance of $\alpha(t,\omega)$ over these windows by:

$$\overline{\alpha}(\Gamma_k,\omega_j) = \frac{1}{M}\sum_{i=1}^{M}[\alpha(t_i,\omega_j)] \quad (11)$$

$$var[\alpha(t,\omega_j)]_{(\Gamma_k,\omega_j)} = \frac{1}{M}\sum_{i=1}^{M}[(\alpha(t_i,\omega_j)-\overline{\alpha}(\Gamma_k,\omega_j))^2] \quad (12)$$

If e.g. $S_2(t_i,\omega_j)=0$ for these M windows, then Equ. (9) shows that $\alpha(t_i,\omega_j)$ is constant over them so that its variance is equal to zero. Conversely, if both $S_1(t_i,\omega_j)$ and $S_2(t_i,\omega_j)$ are different from zero *AND* non constant values over (Γ_k,ω_j), then $var[\alpha(t,\omega_j)]_{(\Gamma_k,\omega_j)}$ is significantly different from zero.
So by searching for the lowest value of expression (12) vs all the available series of windows (Γ_k,ω_j), we directly find a time-frequency domain (Γ_k,ω_j) where only one source is present. The corresponding value of c to cancel this source is then given by the mean computed in Equ. (11).
To find the second separating coefficient, we just have to check the next lowest value of expression (12) vs (Γ_k,ω_j) which gives a significantly different c. A difference of 10^{-1} is a good practical value, allowing hard mixtures, where both separating coefficients c_1 and c_2 are of similar range.
We now have the two best estimated values of the correct separating coefficients given in Equ. (3).

4. EXTENSION TO THE UNDERDETERMINED CASE

The previous criterion allows one to cancel one source in the observations if there exists a time-frequency window where only this source occurs. This criterion may be extended to the case when we have 2 observations of N sources.
In this case, the observed signals become:

$$\begin{cases} x_1(t)=\sum_{m=1}^{N} a_{1m}s_m(t) \\ x_2(t)=\sum_{n=1}^{N} a_{2n}s_n(t) \end{cases} \quad (13)$$

The complex ratio $\alpha(t,\omega)$ of Equ. (8) here reads:

$$\alpha(t,\omega)=\frac{\sum_{m=1}^{N} a_{1m}S_m(t,\omega)}{\sum_{n=1}^{N} a_{2n}S_n(t,\omega)} \quad (14)$$

One can see in Equ. (14) that if only source k exists in a time-frequency window (t_i,ω_j) we have exactly the same expression as in Equ. (10), i.e:

$$\alpha(t_i,\omega_j)=\frac{a_{1k}}{a_{2k}} \quad (15)$$

This value gives the exact coefficient to cancel the contribution from source $s_k(t)$ in the observations by using (2). The only restriction is, once again, that there must exist a time-frequency window where only this source occurs and that the time-frequency transform of each source is not constant over (Γ_k,ω_j).
This solution is perfectly suited to noise reduction for example. By determining a time-frequency window where only the noise occurs this method gives an efficient solution to cancel it, under the assumption that the source signal considered as noise is the same in both observations, up to a scale factor.
This method also applies to karaoke-like applications. Using the stereo observation of a recorded song, we are able under assumption 1. and 2. to cancel the contribution of a singer or an instrument. This performs perfect source cancellation if no global stereo reverberation is added in the song, which would transform the instantaneous mixture in a convolutive mixture[3]. Moreover, experimental tests show that even in this latter case we cancel an important part of one source because the reverberation normally has a lower level than the instantaneous contribution. The main drawback for such applications is that the linear combination between the observations performed by our method, as shown in (2), changes the "balance" between the instruments and gives a "mono" output.

5. EXPERIMENTAL RESULTS

5.1. Configuration with two mixtures of two sources

We choose the mixing matrix as:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0.9 \\ -0.8 & 1 \end{bmatrix} \quad (16)$$

The two theoretical separating coefficients are, according to Equ. (3): $c_1=0.9$ and $c_2=-1.25$.

This first test has been performed using two different voice signals recorded from the radio at a sampling rate of 8000 Hz. We compute the short time Fourier Transform on 128-sample half-overlapping windows, which equates to 16 ms. The time period Γ_k for variance analysis consists of $M=10$ of these windows, which means that a source is only requested to occur alone in one frequency window during 160 ms to be cancelled. With these settings our method yields $c_1=0.8999$ and $c_2=-1.2508$, which is quite close to the target values. Respective observed variances are 2.0651e-4 and 5.4819e-4. Figures 1 to 6 show the temporal representation of the sources, mixtures and output signals. Figures 7 to 10 show the time-frequency analysis of these sources and mixtures signals. One can see that the time-frequency representations of the sources in Figures 7 and 8 are slightly different. These signals can be considered

[3]Usually, all the instruments are recorded one by one and then artificially mixed using linear instantaneous mixing devices.

as a "difficult configuration" because the formants of both voices are present in nearly the same time-frequency areas. The two mixtures in Figures 9 and 10 are very similar and the plain ratio $\alpha(t,\omega)$ shown in Figure 11 does not allow one to localize the constant values domains,which shows the need to compute the variance of this ratio as described in (12). For better legibility the inverse of the variance is presented in Figure 12. This representation enhances the domains where the variance is low. One can easily see which time-frequency domains provide the proper solutions for the separating coefficients. Figure 5 and 6 show that the separation is achieved with high resolution. On listening to these signals the difference between the original and separated signals is not perceptible.

5.2. Configuration with two mixtures of three sources

We recorded a stereo song with continuous voice and two guitars which play nearly the same instrumental part. The purpose here is to show the ability of the proposed approach to cancel the voice from the mixtures, although the guitars are continuously playing. All these sources were recorded one by one on a 4-track magneto recorder with a SNR around 60 dB. We sampled the signals from the console at 44100 kHz with 16-bit resolution and then artificially mixed them with the following mixing matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.4 & 0.8 \\ 0.3 & 0.8 & 0.8 \end{bmatrix} \quad (17)$$

a_{13} and a_{23} are for the voice whereas the other coefficients are for the guitars. We chose to put the voice in the middle of the stereo, like in a regular mix. Thus the theoretical separating coefficient for the voice is $c_3 = 1$. With Equ. (15) one can also see that the two separating coefficients which allow to separately cancel each guitar are : $c_1 = 0.5$ and $c_2 = 2.33$. The length of the time windows for Fourier transform is set to 256 samples, which corresponds to 5.8 ms. The variance is then computed on 10 of these windows, which is equal to 58 ms. We used 4.3 seconds of the song when all three sources are present to compute the separating coefficients. We used the method proposed in the previous subsection and we obtained the two separating coefficients $c_1 = 1.00003234$ with a variance of 2.0466e-8 and $c_2 = 0.5136$ with a variance of 2.3421e-4. Thus the voice cancellation is nearly perfect. The obtained output is a mono signal with the new mixing coefficients, given by (17) and (2): $out = 0.399990298 * (\text{guitar1}) - 0.400025872 * (\text{guitar2}) - 2.5872e-5 * (\text{voice})$ which gives an ideal karaoke playback. As we have nearly half the power of guitar in the output as compared to any input, an approximate value of the voice attenuation is given by $20 * log(\frac{2*2.5872e-5}{.8}) = -193dB$. Figures 13, 14 and 15 show the time frequency representations of the first, second guitar and the voice. One can see that one of the two guitars contains more frequency component than the other one which is confirmed by listening to their respective sounds. Thus even if both guitars play the same instrumental part, there exist some differences in the time-frequency representations of their signals. We notice in Figure 15 that unlike guitars, the voice includes high-medium and high frequency component which are situated between 7 kHz and 15 kHz. Thus only the voice exists in this frequency band. None of these three sources contains high frequency above 15 kHz. So, the remaining signal between 15 kHz and 22 kHz is some noise. Figures 16 and 17 show the time-frequency representation of the left and right sides of the stereo input which look very similar. The inverse variance graph in Figure 18 is interesting. We can see on it that most low-variance points are in a frequency band, i.e. 7 to 15 kHz, where only the voice is present. No low-variance point exists for frequencies higher than 15 kHz because no source occurs in these regions and the respective noises added to each source do not produce constant time-frequency values, i.e. are not short-time stationary. Only few low variance points exist for frequencies lower than 7 kHz because both guitars occur, play the same chords and the voice has the same fundamental tone. So it is hard to find some time-frequency areas with only one source below 7 kHz. Our method performs voice cancellation by self-focusing on the time-frequency frequency domains where only the voice is present. It also gives a separating coefficient to cancel a guitar, which might be hard to find because of the similarity of the produced sounds.

We demonstrated here that the time-frequency information allows to perform a nearly perfect source cancellation. We obtained similar results on mixtures realised on a "studio mixing console".

6. CONCLUSION

We proposed here an efficient method for solving the linear instantaneous blind source separation problem with mixtures of 2 sources. This method also performs very well in karaoke-like applications when only two observations of more than two sources are available.

Unlike classical methods [1], this new approach based on time-frequency analysis only needs the sources to be non-stationary and to have some differences in their time-frequency representations. Thus no assumption is made about the gaussianity, coloration or independence of the sources. This allows one to separate some signals which are often excluded from other methods. Moreover this method directly achieves source cancellation without any convergence issues and is much simpler than the few time-frequency methods that were previously reported [6], [7]. Many tests have been performed on speech or music samples and show the

robustness of this approach.

7. REFERENCES

[1] J. F. Cardoso, "Blind signal separation: statistical principles," in *Proceedings of the IEEE*, October 1998, vol. 86, number 10, pp. 2009–2025.

[2] Y. Deville, "A source separation criterion based on signed normalized kurtosis," in *Proceedings of the 4th International Workshop on Electronics, Control, Measurement and Signals (ECMS'99)*, Liberec, Czech Republic, May 31 - June 1, 1999, pp. 143–146.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Zibulevsky and B. A. Pearlmutter, Blind source separation by sparse decomposition in a signal dictionary, in *Independent component analysis: Principles and practice*, Robert S. J. and Everson R. M. editors, Cambridge University Press, 2000.

[8] J. K. Hammond and P. R. White, "The analysis of non-stationary signals using time-frequency methods," *Journal of sound and vibrations*, pp. 419–447, 1996.

[9] 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.

[10] L. Cohen, *Time-frequency analysis*, Prentice hall PTR, Englewood Cliffs, New Jersey, 1995.

[11] L. Cohen, "Time-frequency distributions - a review," in *Proceedings of the IEEE*, July 1989, vol. 77, No. 7, pp. 941–979.

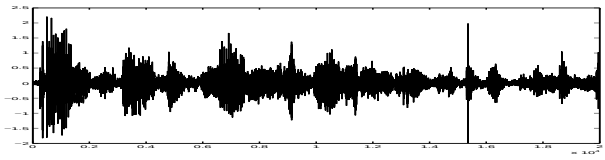

Fig. 1. Source s_1 in time domain

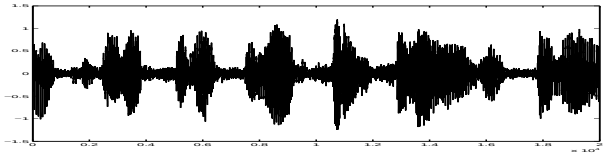

Fig. 2. Source s_2 in time domain

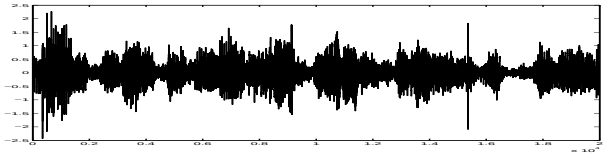

Fig. 3. Mixed signal x_1 in time domain

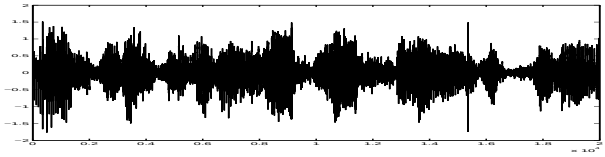

Fig. 4. Mixed signal x_2 in time domain

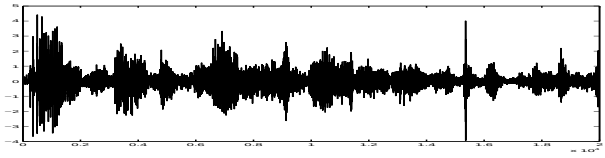

Fig. 5. Output signal y_1 in time domain

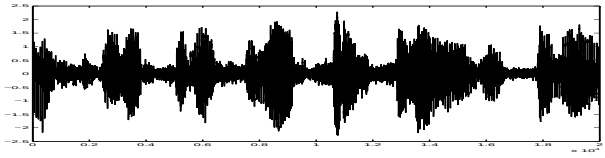

Fig. 6. Output signal y_2 in time domain

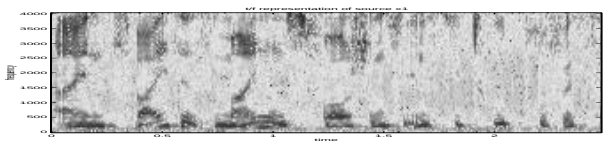

Fig. 7. Time Frequency representation of s_1

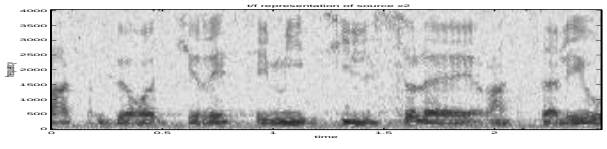

Fig. 8. Time Frequency representation of s_2

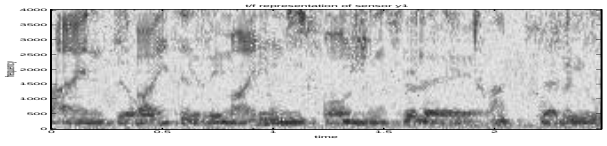

Fig. 9. Time Frequency representation of x_1

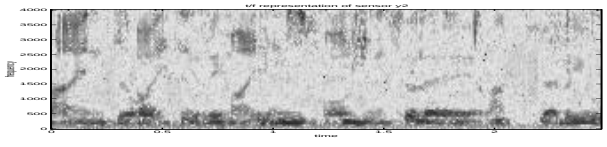

Fig. 10. Time Frequency representation of x_2

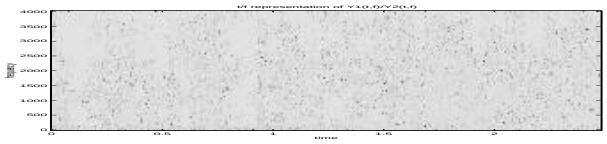

Fig. 11. Time Frequency representation of $X_1(t,\omega)/X_2(t,\omega)$

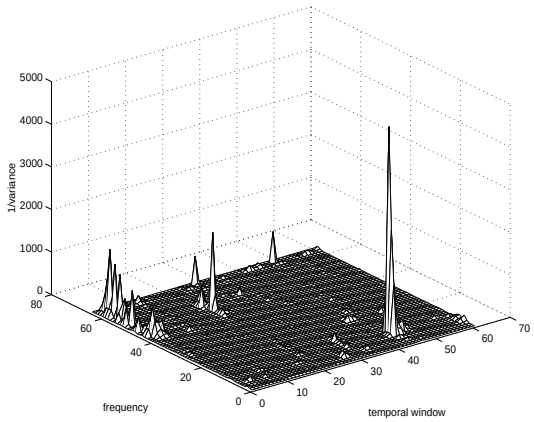

Fig. 12. Time Frequency representation of $\frac{1}{var[X_1(t,\omega)/X_2(t,\omega)]}$. Axes units: windows indices

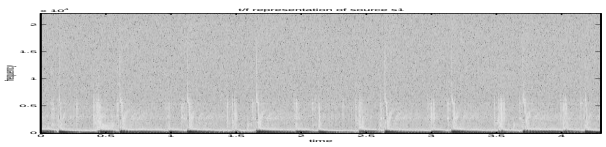

Fig. 13. Time Frequency representation of guitar s_1

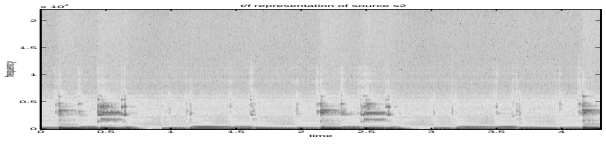

Fig. 14. Time Frequency representation of guitar s_2

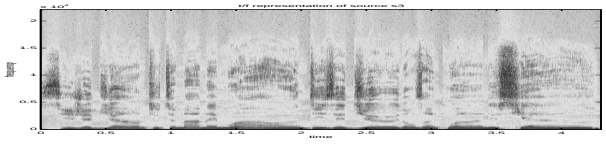

Fig. 15. Time Frequency representation of voice s_3

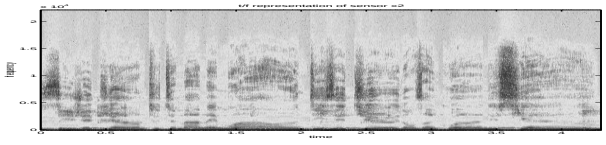

Fig. 16. Time Frequency representation of x_1

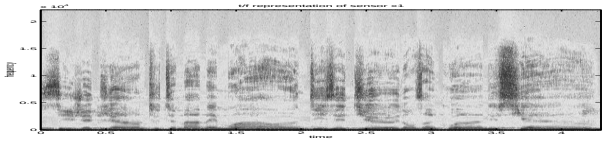

Fig. 17. Time Frequency representation of x_2

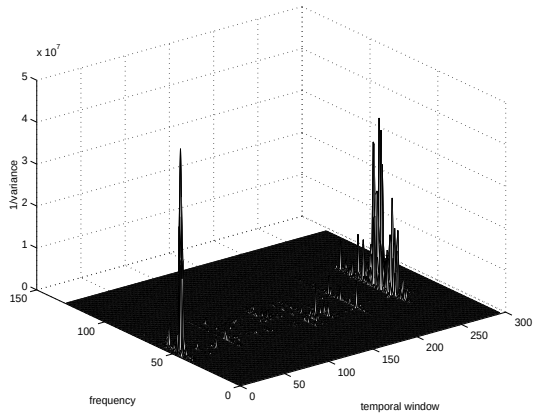

Fig. 18. Time Frequency representation of $\frac{1}{var[X_1(t,\omega)/X_2(t,\omega)]}$. Axes units: windows indices