5th IEEE Hiroshima Student Symposium, December 2003, Hiroshima, Japan

A Method for Segregation of Speech Signals

E. MOLINE[1](Undergraduate student in ENSIETA, Brest – FRANCE)
J. TSUTSUI (Master Student, Dept. of Electronic and Control Systems Eng., Shimane University, JAPAN)
T. OKAMOTO (Master Student, Dept. of Electronic and Control Systems Eng., Shimane University, JAPAN)
Ali MANSOUR[2] (Lab. E3I2, ENSIETA, Brest – France)
M. KAWAMOTO[3] (Dept. of Electronic and Control Systems Eng., Shimane University, Shimane, JAPAN)
Y. INOUYE (Dept. of Electronic and Control Systems Eng., Shimane University, Shimane, JAPAN)

Emails : [1]molineer@ensieta.fr, [2]mansour@ieee.org, [3]kawa@ecs.shimane-u.ac.jp,

ABSTRACT
This paper deals with the problem of blind separation of under-determined or over-completes mixtures (i.e. more sources than sensor) and more particularly in the special case of two speech signals and one sensor. After a introduction to the problem of blind separation of sources, an algorithm based on time-frequency representations (TFR) is presented. Finally, some experiments are conducted and some experimental results are given.

1. INTRODUCTION

The problem of blind separation of sources (BSS) is a recent and important signal processing problem. This problem involves the finding of unknown sources only by observing some mixed signals of them. This problem was first put forward by Hérault *et al* [1] in biology field. Biological sensors are sensitive to many sources, therefore the central nervous system processes of typical multidimensional signals. So BSS was proposed as a mathematical approach to mimic a biological system.

If H is the transfer function between source signals and sensors (which depends only on the channel and the sensors parameters), the separation consists on the estimation of this unknown transfer function H (or its inverse) by only using the observed signals. Conventionally, researchers consider two main linear models of transfer functions: a memory-free channel (instantaneous mixture) and a memory channel (convolutive mixture). The following assumptions are widely used. First, the sources are statistically independent with each other. Second, the number of sensors should be greater than or at least equal to the number of sources (for subspace approaches). Finally the sources have a non-Gaussian distribution, or more precisely, at most one of them can be a Gaussian signal. Under these mild assumptions, P. Comon proved that the blind separation is possible and he proposed a theoretical concept named ICA (i.e. Independent component analysis), see [2]. Recently, many researchers proposed ICA algorithms to deal with application fields [3], [4] such as: the separation of biomedical signals, for example the separation of the heartbeat electrocardiography of a mother and the heartbeat ECG of her fetus [5]. Or also in radio communication fields, especially for mobile-phones (SDMA, Spatial Division Multiple Access) or free-hand phone applications [6]. It is also used in some visual image communication system [7], to separate seismic signals [8] and for airport surveillance [9], [10]. In these diverse applications the last three assumptions are usually fulfilled. Nevertheless in recent applications (e.g., robots which imitate human behavior, double-talk in satellite communication, biomedical applications as EEG or MRI signals) the number of sources is greater than the sensors which are often mono-sensor.

Recently, to separate under-determined or over-completes mixtures, a new approach based on time-frequency representation (TFR) has been proposed [11]. The major drawbacks of that algorithm were the used of a quadratic non-invertible transformation (the Pseudo-Wigner-Ville) and the difficulties to make filters in TFR domain. To improve the previous algorithm, we are suggesting here a linear TFR such Short Fast Fourier Transform. Using the spectrogram of the signals we propose in this manuscript a new algorithm.

2. A TIME-FREQUENCY APPOARCH

The algorithm proposed in this section is based on time-frequency distributions of the observed signal (TFR) to separate two speech sources. The TFR is a 2D representation of a signal. Time is on the X-axis, frequency is on the Y-axis and the power value of the frequency is represented by a gray scale. It is possible to detect the energetic areas of the signal. It is known that speech signals are non-stationary signals. However within phonemes (duration of about 70 ms) the signal's statistics are relatively constant. Moreover it is well

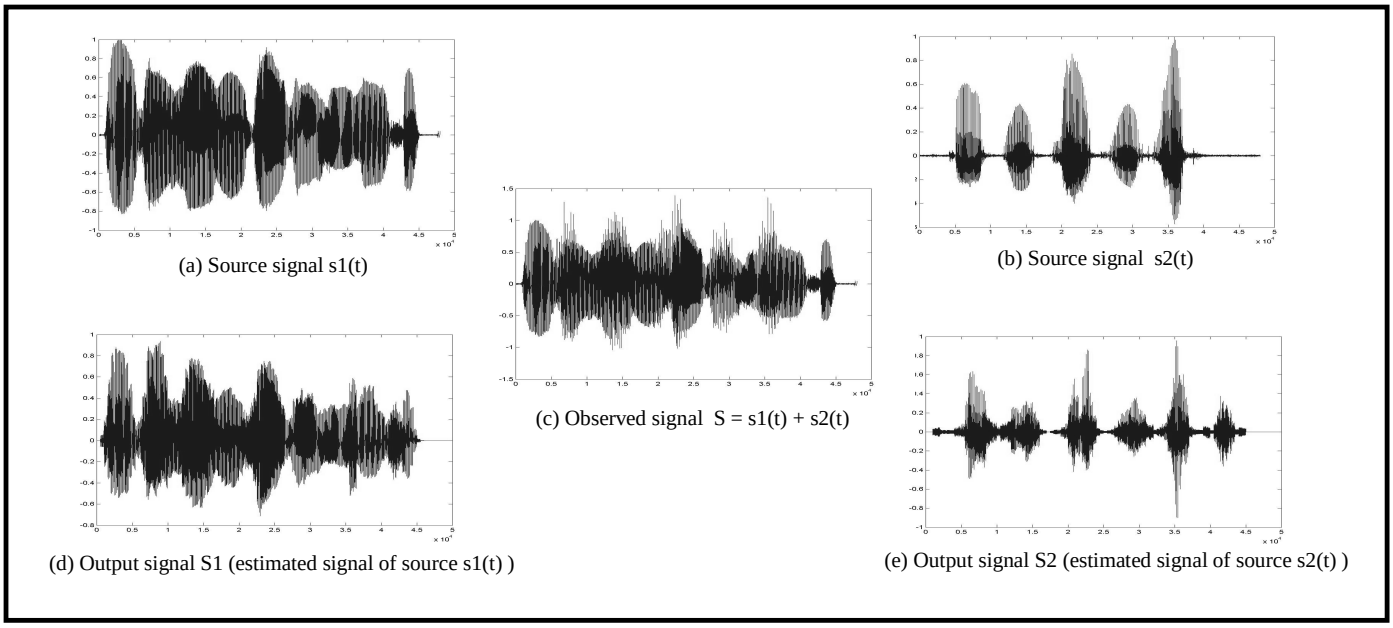
known that voiced speech is a quasi-periodic signal. The main idea of the proposed algorithm is based on the existence of a pitch which can characterize a speaker (and therefore a signal). This pitch is linked with harmonics frequencies. The pitch and the formants (resonance of the buccal cavity) correspond to the more energetic area of the spectrogram. So our goal is to detect these different frequencies in Time-Frequency Domain and to gather each harmonics frequency with the related pitch in order to create, in Time-Frequency domain a filter used to create the estimated signals. A strong assumption is also made: the fundamental frequencies of the two speakers must be very different.

At first one should calculate the TFR of the observed signal. This is done using a Hamming window which measures about 20 ms. This time value is a good estimation of the time spent to say a basic sound. At every step**,** the Hamming window is applied to the signal and a Fast Fourier Transform (FFT) is calculated to obtain the frequency representation of the selected part of the signal. Then the Hamming window is shifted to analyze the next part of the signal. The multiplication between the signal and the window can be regarded as a weighting. Therefore there is a loss of information. So the best way to correctly rebuild the signal later is to shift the Hamming window by half of its length. To obtain a more precise and useful TFR, one first uses a high-pass filter in order to increase the high frequencies which are less powerful in a human voice and also to reduce interference. In this way one can obtain a TFR with a better resolution. Another classical technique used to improve the resolution in the TFR is zero-padding. Finally to obtain an easily read spectrogram, all low power frequencies (under -3 dB) are not selected.

Later on, in the time-frequency domain we check every Δt time (about 10 ms, half the size of the Hamming window) for the possible peaks of frequency. For every Δt time slice a frequency list is made, which is called a "detection list". This is a list of frequencies which includes all the harmonics frequencies, the fundamental frequencies and even some frequencies due to noise. The pitch of a human voice is a relatively low frequency and it depends on the physical characteristics of the speaker (size and form of the buccal cavity) and on his mood (angry or sad) or health (a cold). But an average man's voice is around 100-150 Hz, a woman's voice is around 140-250 Hz and a child's voice is around 250-400 Hz. After the detection, the different frequencies are sorted by increasing order. Then, one can assume that the lower frequency which is higher than 70 Hz (under this value it is noise) becomes a pitch. At this stage, the obtained different frequencies are classified into two sets: a multiple of the selected pitch (relative error less than 5 %) and the others. Finally a "signal_1 list" which contains the frequencies of one signal and a "rest list" which contains the other frequencies are obtained. This process should be repeated with the non-selected frequencies, should be repeated until the "signal_2 list" is obtained. Indeed, because the frequencies of the "signal_1" is removed from the "detection list", the first frequency that we come across, which is less than 500 Hz, could be the pitch of signal 2. Nevertheless it is very important to note that the first frequencies in the two Δt times are not necessarily the pitch of the same signal. Because one source might be silent over a short time. Therefore since the obtained two lists may contain errors, we need to correct. This is easily done by comparing the pitch to a threshold**.** This threshold can be fixed, based on a histogram of the different pitch values. In most cases, one can see that the frequencies are located in two main areas (the pitch of the speaker 1 and of the speaker 2) and the threshold can be chosen between these two areas. If the time-frequency signatures of the two sources are not disjoint, then the pitches are too close. In the present paper, however, we assume that the fundamental frequencies of two speakers are very different. Finally, we obtain two matrices (signal_1 and signal_2 list for each Δt time slice) which are used to filter the matrix of the spectrogram. After having selected the desired frequencies in the spectrogram one can obtain the two estimated signals by coming back to the time domain. Many estimated time slices of the signals are obtained by using an Inverse Fast Fourier Transform at every Δt time slice. Using some shifting techniques, the different slices are correctly merged into an estimated signal.

3. EXPERIMENTAL RESLUTS

In this section we assume that the sources have different time-frequency signatures. If this assumption cannot be satisfied, the classification part of the algorithm cannot be successfully achieved. To validate the proposed algorithm some computer experiments were conducted. Several male voice sources were recorded with a 16 KHz sampling frequency for 3 seconds. Well-pronounced, but rather slow sentences were recorded. The TFR was calculated by using 1000 samples of the observed signal. Figure 1 and figure 2 show the results obtained by applying the algorithm described in section 2 to the observed signal which is a sum of two male voices. From this figure, one can see that the first results are very promising. Moreover, if one listens to the results, it can be heard that the estimated signals are very close to the original ones.



(a) Source signal s1(t)

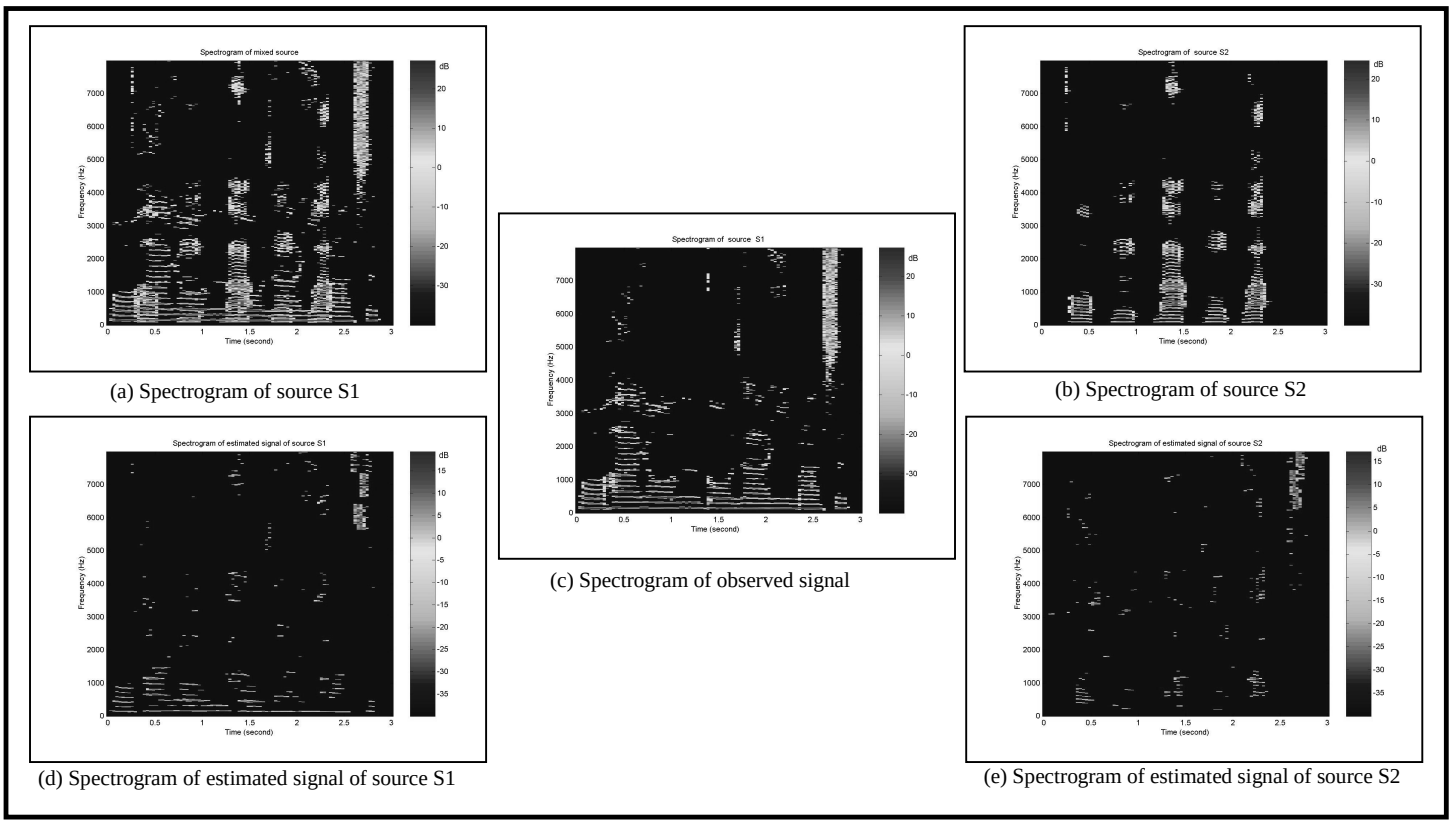
(b) Source signal s2(t)

(c) Observed signal S = s1(t) + s2(t)

(d) Output signal S1 (estimated signal of source s1(t))

(e) Output signal S2 (estimated signal of source s2(t))

Fig 1 Simulations Results



(a) Spectrogram of source S1

(b) Spectrogram of source S2

(c) Spectrogram of observed signal

(d) Spectrogram of estimated signal of source S1

(e) Spectrogram of estimated signal of source S2

Fig 2 Simulations Results

4. CONCLUSION

This paper deals with the problem of blind separation of under-determined (or over-complete) mixtures and more precisely in the case of two human voice sources observed with one sensor. A segregation algorithm based on the Time-Frequency Representation and using human voice properties has been proposed. Then some experiments were carried out and some results produced to validate it. Finally let's put forward some improvements which we are currently working on.

A future work will be conducted to improve the performance of the proposed algorithm. It is clear that the choice of the TFR is a crucial one. On the other hand the actual version of the algorithm is using the Short Fast Fourier Transform SFFT and the Hamming window which is a classic and standard TFR. Therefore, different and more modern adequate TFR should be considered. Of course a second improvement could be the detection-classification phase. For instance, in the proposed algorithm the segregation only takes into account some properties of the human voice, such as the quasi-periodic and the pitch or the fact that even if speech signals are non-stationary signals they are relatively constant within phonemes. To improve the detection-classification stage one should integrate some correlation procedures to better merge the different obtained time slices into the wave voice of one person. Finally one can see if it is better to create a filter in time-frequency domain and then use an Inverse Fast Fourier Transform, or if it is better to use information obtained in TFR to create filters which are directly applied to the observed signal.

5. REFERENCES

[1] J. Hérault, C. Jutten and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé" in *Actes du Xème colloque GRETSI*, pp. 1017-1022, Nice, France, 20-24, Mai 1985.

[2] P. Comon, "Independent component analysis, a new concept?" , pp. 287-314, April 1994.

[3] A. Mansour, A. Kardec Barros, and N. Ohnishi "Blind Separation of Sources: Methods, Assumptions and Applications." IEICE Trans. Fundamentals, vol.E83-A, no. 8, pp. 1498-1512, August 2000.

[4] A. Mansour and M. Kawamoto "ICA papers classified according to their applications and performances" IEICE Trans. Fundamentals, vol.E86-A, no. 3, pp. 620-633, March 2003.

[5] L. De Lathauwer, D. Callaerts, B. De Moor, and J. Vandewalle, "Separation of wide band sources" *in HOS 95*, pp 134-138, Girona-Spain, 12-14 June 1995.

[6] N. Charkani, "Séparation auto-adaptative de sources pour les mélanges convolutifs. Application à la téléphonie mains-libres dans les voitures" Ph.D. thesis, INP Grenoble, Novembre 1996.

[7] H. Szu, T. Yamakawa, and C. Hsu, "Visual image communication using advanced neural networks" in *First International Workshop on Independent Component Analysis and signal Separation (ICA99)*, pp 121-126, Aussois, France, 11-15 January 1999.

[8] N. Thirion, J. MARS, and J. L. BOELLE, "Separation of seismic signals: A new concept based on a blind algorithm" in *Signal Processing VIII, Theories and Application*, pp. 85-88, Elsevier, Triest, Italy, September 1996.

[9] G. Desodt and D. Muller, "Complex independent components analysis applied to the separation of radar signals" in *Signal Processing V, Theories and Applications*, L. Torres, E. Masgrau, and M. A. Lagunas, Eds. , pp. 665-668, Elsevier, Barcelona, Spain, 1994.

[10] E. Chaumette, P. Common, and D. Muller, "Application of ICA to airport surveillance" in *HOS 93*, pp. 210-214, South Lake Tahoe-California, 7-9 June 1993.

[11] A. Mansour, M. Kawamoto, C. Puntonet, "A time-frequency approach to blind separation of under-determined mixture of sources" Proceedings of the IASTED International Conference on Applied Simulation and Modeling (ASM2003), Marbella, Spain, 3-5 September 2003.

[12] P. ALLUIN, "Analyse et Traitement du signal acoustique"

[13] T. DUTOIT, "Un bilan des développements récents en traitement automatique de la parole", Faculté Polytechnique de Mons, 2000.

[14] L. BUNIET, "Traitement automatique de la parole en milieu bruité: étude de modèles connexionnistes statiques et dynamiques", février 1997.

[15] Y. LAPRIE "Analyse spectrale de la parole", octobre 2000.