Convolutive Blind Source Separation with Wiener Post-Filtering for Robust Speech Recognition*

Leandro Di Persia[1,2], Diego Milone[1,2], and Masuzo Yanagida[3]

[1] Grupo de Investigación en Señales e Inteligencia Computacional. Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Argentina
[2] Laboratorio de Cibernética. Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
[3] Department of Knowldge Engineering, Doshisha University, Japan
ldipersia@gmail.com, d.milone@ieee.org,
myanagid@mail.doshisha.ac.jp

Abstract. Blind source separation for convolutive mixtures of sound sources is a complex task, mainly because the mixing filters are long and non-minimum phase. One approach to solve this problem is frequency domain blind source separation, in which the separation is calculated for each frequency bin in the time-frequency domain. Although there are several methods for this task, separation quality is degraded by many factors. This paper presents a method for separation in time-frequency domain, that combines the advantages of other two separation methods and uses a time-frequency Wiener filter as post-processing to increase separation quality. The algorithm has been evaluated over a database of Spanish speech recorded in a reverberant room using two active sound sources and two microphones. Speech recognition results show an increment in recognition rate of the separated speech in the order of 70% from the noisy case.

1 Introduction

The objective of blind source separation (BSS) consist of, given a set of sound field measurements obtained by means of microphones in specified locations, to obtain a set of signals approximating the original sound sources that have produced the sound field. In the case of free-field propagation of sound (i.e. in open spaces without enclosures), the sound wave originated in each source arrives only one time to each sensor. In such a way, the mixture can be considered as a linear additive mixture. On the contrary, when the mixture is produced inside an enclosed environment, the sound waves are reflected by every solid surface in the room and so each microphone receives not only the direct sound wave but also all the reflections, and more over, the reflections of all orders until energy of

* This work is supported by ANPCyT-UNER, under Project PICT N 11-12700, UNL-CAID 012-72 and CONICET

the source vanishes. This phenomenon, called reverberation, can be modeled as the output of an LTI system [1], that is, as a convolution between the original sound source and the impulse response of the room.

As a result of reverberation, the mixture as recorded at microphones is not simply additive, but it must be considered as a convolutive mixing, in such a way that each microphone is excited by the addition of filtered versions of the original sources. This reverberation phenomenon produces echoes and spectral distortion that degrades recognition rates in case of automatic speech recognition (ASR) systems [2], even if the system is trained with reverberant signals recorded in the same room [3].

Given a number M of active sources and a number N of sensors, with $N \geq M$, assuming that the environment effect can be modeled as the output of an LTI system, measured signals at each microphone can be modeled as a convolutive mixture model [4]:

$$x_j(t) = \sum_{i=1}^{M} h_{ji}(t) * s_i(t) \tag{1}$$

where x_j is the j-th microphone signal, s_i is the i-th source, h_{ji} is the impulse response of the room from source i to microphone j, and $*$ stands for convolution. This equation can be written in compact form as:

$$\mathbf{x}(t) = \mathbf{H}(t) * \mathbf{s}(t) \quad . \tag{2}$$

Taking a short-time Fourier transform (STFT) of the previous equation, the convolution becomes a multiplication, and assuming that the mixture filters are constant over time (that is, impulse responses does not vary in time), this can be written as:

$$\mathbf{x}(\omega, \tau) = \mathbf{H}(\omega)\mathbf{s}(\omega, \tau) \quad . \tag{3}$$

Thus, for a fixed frequency bin ω this means that a simpler instantaneous mixture model can be applied. Under the assumption of statistical independence of the sources over the STFT time τ, the separation model for each frequency bin can be solved using one of the methods for Independent Component Analysis (ICA) [5]. In this context, for each frequency bin ω a matrix $\mathbf{W}(\omega)$ is searched such as:

$$\mathbf{y}(\omega, \tau) = \mathbf{W}(\omega)\mathbf{x}(\omega, \tau) \tag{4}$$

where resulting separated bins $\mathbf{y}(\omega, \tau)$ should be approximately equal to the original $\mathbf{s}(\omega, \tau)$.

To simplify notation, as from now on all equations would be dealing with time-frequency representations, we will obviate time and frequency variables, and so, for example, $\mathbf{x}$ must be interpreted as $\mathbf{x}(\omega, \tau)$, except if the context makes confuse that interpretation, in such case it will be explicitly written.

To estimate separation matrix $\mathbf{W}$ several algorithms have been applied. Some authors have used second-order statistics and decorrelation procedures [6,7,8].

Others have proposed the use of fixed-point algorithms derived from FastICA algorithm [9,10,11]. Some information theory derived algorithms based on minimization of mutual information [12], information maximization (InfoMax) [13] or Kullback-Leibler divergency [14], combined with Natural Gradient [4] have been also successfully used. Recently also, some algorithms combining several techniques have been presented, as in [15] where a combination of FastICA followed by an ICA by InfoMax with Natural Gradient is used.

In this paper the separation is solved by a combination of two methods. Then, a Wiener time-frequency filter is estimated and applied in order to improve the separation. In the following sections, the algorithm will be explained in detail. Then, some experiments to evaluate the quality of the separation will be presented, followed by results, analysis of results and finally some conclusions and future works.

2 Separation Algorithms

For each frequency bin, two separation algorithms for complex-valued signals are sequentially applied. In the first stage, Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm [16] is applied to obtain a first estimation of separation matrix $\mathbf{W}$. Then, this separation matrix is refined by using it as initial condition for FastICA algorithm [17].

After separation in each frequency bin there are two problems to solve. All ICA algorithms are able to obtain an estimation of sources up to an scaling and permutation indeterminacy, what means that for each frequency the resulting sources will have different scaling and different sorting. Thus, before any other step, one need to order the sources and obtain a consistent scaling for each frequency.

Following this, a Wiener time-frequency filter estimated from the separated sources is applied. Finally an inverse STFT is applied to yield time-domain sources. In the following subsections, these aspects will be discussed in detail.

2.1 JADE Algorithm

JADE algorithm is an independent component analysis (ICA) method that uses explicit High Order Statistics by means of fourth-order circular cumulant tensor. For a random vector X with probability density function $f_X(\mathbf{x})$, the fourth order circular cumulant is given by:

$$C^{jl}_{ik,X} = Cum\left\{X_i, X_j^*, X_k, X_l^*\right\} \quad . \tag{5}$$

Cardoso et al [16] proposed to obtain the unitary matrix $\mathbf{W}$ by means of maximizing the cost function

$$\mathcal{J}(\mathbf{W}) = \sum_{i,k,l=1}^{M} \left|C^{il}_{ik,Y}\right|^2 \tag{6}$$

where $\mathbf{y}$ is observed as in (4). This optimization is equivalent to a join diagonalization of a set of eigen-matrices.

Given a matrix P, the fourth-order cumulant $C^{jl}_{ik,X}$ defines a linear transformation $\Omega(P)$ in such a way that

$$\Omega(P)_{i,j} = \sum_{k,l} C^{jl}_{ik,X} P_{k,l} \ . \tag{7}$$

This linear transformation has M^2 eigen-matrices $E_{p,q}$ that satisfy $\Omega(E_{p,q}) = \lambda_{p,q} E_{p,q}$. It is enough to find the M most significant eigenmatrices (i.e. those with bigger associated eigenvalues) and perform approximate joint-diagonalization of these eigen-matrices to obtain separation matrix $\mathbf{W}$ for signals in $\mathbf{x}$. Before applying this algorithm, a whitening transformation is performed in order to eliminate second order correlations and simplify the algorithm convergence.

2.2 FastICA Algorithm

Although JADE has good separation capabilities, its performance can be improved by using the separation matrix obtained as initial value for another algorithm that refines it by optimizing some contrast function. In this case, FastICA algorithm has been used [17].

This algorithm uses a deflationary approach where each source is extracted sequentially. Therefore, for each source a separation vector of signal in frequency domain $\mathbf{w}_i$ is pursued such that $\tilde{s}_i = \mathbf{w}_i^H \mathbf{x}$ will be approximately one of the sources. To search for the proper vector $\mathbf{w}_i$, an optimization problem is solved. This optimization is stated as maximizing

$$\sum_{i=1}^{M} J_G(\mathbf{w}_i) = \sum_{i=1}^{M} E\left\{G\left(\mathbf{w}_i^H \mathbf{x}\right)\right\} \text{ with respect to } \mathbf{w}_i, \ \ i = 1, \ldots, M \tag{8}$$

subject to

$$E\left\{\left(\mathbf{w}_i^H \mathbf{x}\right)\left(\mathbf{w}_j^H \mathbf{x}\right)\right\} = \delta_{ij} \ . \tag{9}$$

In this equations, $E\{\cdot\}$ is the expectation operator, and $G : \mathbb{R}^+ \cup 0 \rightarrow \mathbb{R}$ is a smooth even function. For this work we have used function $G(y) = \log(\alpha + y)$, with $\alpha = 0.1$.

To achieve this, the contrast function $J_G(\mathbf{w}_i)$ is maximized to obtain a separation vector $\mathbf{w}_i$, then a deflationary Gram-Schmidt-like decorrelation is used to eliminate the information of previously obtained sources and this process is iterated until all desired sources are extracted. It must be noted that matrix $\mathbf{W}$ will have $\mathbf{w}_i^H$ as its i-th row. An alternative to this sequential extraction is to extract all sources at once, optimizing a matrix $\mathbf{W}$ and using an ortonormalization method on that matrix after each iteration. In this paper we have used the deflationary approach.

This is a Newton-like fixed-point iteration with quadratic convergence, and as such, it is very fast. As all fixed point methods, it depends on good initial conditions estimation, and we have a good estimation using the output of JADE algorithm.

2.3 Indeterminacies

To solve scaling and permutation indeterminacies, a variant of the method proposed by [7] has been used. For the scaling ambiguity, the approach consists of recovering the filtered versions of the sources instead of the sources themselves. So, mixtures are modeled as $\mathbf{x} = \mathbf{v}_1, \ldots, \mathbf{v}_M$. Using separation matrix $\mathbf{W}$ and its inverse (i.e. estimated mixing matrix) $\mathbf{W}^{-1}$, one can write:

$$\begin{aligned}\mathbf{x} &= \mathbf{W}^{-1}\mathbf{y} \\ &= \mathbf{W}^{-1}\mathbf{W}\mathbf{x} \\ &= \mathbf{W}^{-1}\mathbf{I}\mathbf{W}\mathbf{x} \\ &= \mathbf{W}^{-1}\left(\mathbf{E}_1 + \cdots + \mathbf{E}_M\right)\mathbf{W}\mathbf{x} \\ &= \mathbf{W}^{-1}\mathbf{E}_1\mathbf{y} + \cdots + \mathbf{W}^{-1}\mathbf{E}_M\mathbf{y} \\ &= \mathbf{v}_1 + \ldots + \mathbf{v}_M \ .\end{aligned} \tag{10}$$

where $\mathbf{E}_i$ is a matrix with a one in the i-th diagonal element and zeros elsewhere. It is easy to prove that the representation of $\mathbf{v}_i$ is independent of the scaling in matrix $\mathbf{W}$.

Now for the permutation problem, the approach makes use of the fact that the envelopes of different sound signals must be different and also, that if the signals are independent the correlation between envelopes of the separated sources must vanish. This must be true for one frequency bin, however one can expect that successive frequency bins should share the same or similar envelopes. This is the information used to solve permutation problem: starting from some frequency band, an estimation of the envelope based on previous classified bands is calculated. Then, for each separated signal in a new frequency bin, correlation between its envelope and the estimated one for pre-classified bins is calculated, and the signal is assigned to that of maximum correlation value [7].

In the original paper, pre-classified envelopes are estimated as an average of all the previously classified envelopes in that class. In this paper, instead of using this approach, we assume that in the averaging process, the last classified envelopes must have more weight since they will be more similar to the envelopes following for classification. Therefore instead of a simple averaging of envelopes, we update that value as

$$\mathcal{E}\left(k\right)_j = \mathcal{E}\left(k-1\right)_j + \alpha E\left(k\right)_j \ . \tag{11}$$

where $\mathcal{E}\left(\cdot\right)_j$ refers to the locally averaged envelope for source class j, and $E\left(\cdot\right)_j$ to the last classified envelope for this class.

After this process we obtain time-frequency representations for each of the source component, in each sensor. That is, for each source we obtain N time-frequency representations, each one corresponding to the effect of that source in one of the sensors, isolated from the other sources effects. Since usually only one representation for each source is needed, in this study we use the alternative of keeping the one with bigger energy to be used in the following steps.

2.4 Time-Frequency Wiener Filter

The separation result will not be perfect mainly due to reverberation times and the simplified time-invariant modeling. When reverberation time increases, the performance of the algorithms tend to decrease. Therefore we propose the use of a non-causal time-frequency Wiener filter as post-processing [18]. Without losing generality, this will be explained for the two sources, two microphones case, and the generalization to more sources being straightforward.

The short-time Wiener filter $H_{\mathcal{W}}$ for a signal generated by the simple additive noise model is:

$$H_{\mathcal{W}}(\omega,\tau) = \frac{\left|\tilde{z}(\omega,\tau)\right|^2 - \left|\tilde{n}(\omega,\tau)\right|^2}{\left|\tilde{z}(\omega,\tau)\right|^2} \tag{12}$$

where $\tilde{n}$ represents the estimated additive noise. In this case we obtain two signals, $\tilde{v}_1$ and $\tilde{v}_2$ and if the separation process was successful one can uses them as estimation of the clean sources.

So, in order to eliminate residual information from source v_2 on source v_1 we can use the short-time power spectrum of $\tilde{v}_1$ as numerator (estimation of clean source) and add short-time power spectrum of $\tilde{v}_1$ and $\tilde{v}_2$ as the estimation of noisy power spectrum in denominator. Moreover, as we know that both signals will have some information from the other, and this sharing would be not uniform over the whole time-frequency plane, one can use time-frequency weights to reduce the effect of the filter, as expressed in the following equation:

$$H_{\mathcal{W},1}(\omega,\tau) = \frac{\left|\tilde{v}_1(\omega,\tau)\right|^2}{\left|\tilde{v}_1(\omega,\tau)\right|^2 + C(\omega,\tau)\left|\tilde{v}_2(\omega,\tau)\right|^2} \tag{13}$$

where the weighting matrix $C\left(\omega,\tau\right) \in [0,1]$.

If the time-frequency contents of $\tilde{v}_1$ and $\tilde{v}_2$ are very similar (so for that time-frequency coordinate the separation was not well done), the weights must be close to zero, otherwise they must be near to one. There are several ways to set these weights. One simple way may include dot products to determine time and frequency similitude of power spectrum.

The short-time Wiener filter to improve source v_2, $H_{\mathcal{W},2}(\omega,\tau)$ is calculated in a similar way to (13), with the roles of v_1 and v_2 interchanged.

3 Results and Discussion

To test the capabilities of this algorithm, some experiments have been made. Sentences for the experiments were extracted from Albayzin Spanish speech corpus [19]. From this big database, a subset of 605 sentences were selected, and those were divided into a training set of 585 a test set of 20 sentences. The training set was used to train a recognizer with clean data. The test set has 5 sentences spoken by 4 speakers. Selected sentences were 3001, 3006, 3010, 3018, 3022, from speakers aagp and algp (female), and mogp and nyge (male). Those 20 sentences were recorded in a room according to Fig. 1.

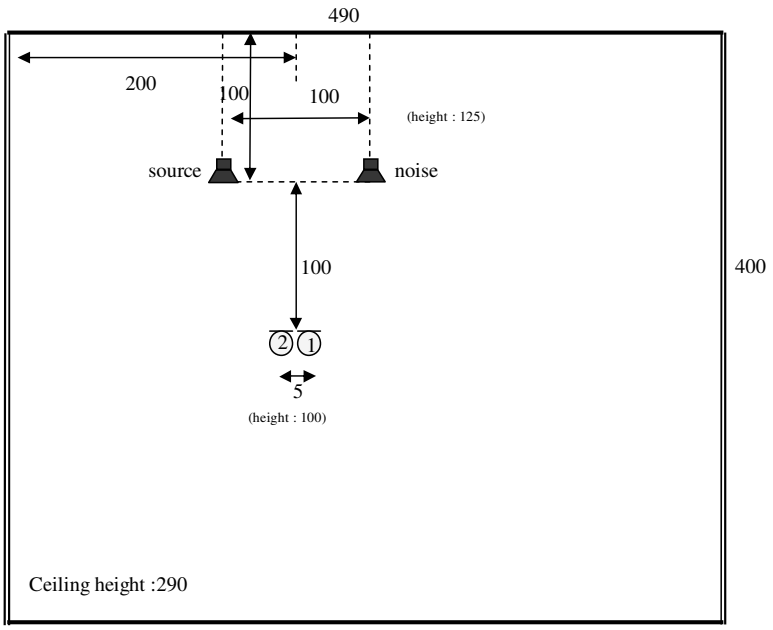

Fig. 1. Experimental setup (all dimensions in cm).

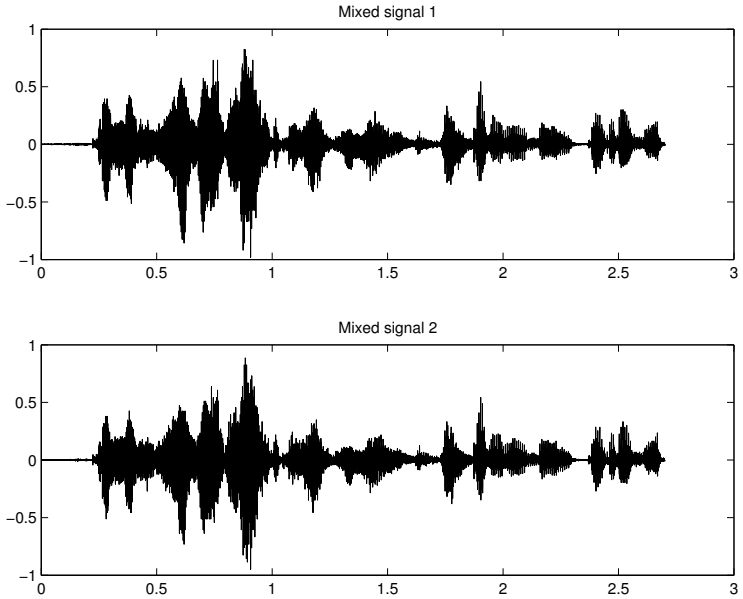

Fig. 2. Mixed signals. Top: microphone 1; bottom: microphone 2. Signal to noise power ratio: 0 dB

Two loudspeakers were used, one to reproduce desired speech source and the other to reproduce some kind of noise. The resulting sound field was recorded with two Ono Sokki MI 1233 omnidirectional measurement microphones, with flat frequency response between 20 Hz to 20 kHz and with preamplifiers Ono Sokki model MI 3110.

Interfering sources were of two kinds, speech and white noise. For speech noise, sentence 3110 and speakers aagp and nyge where selected. To contaminate female utterances, male speech (nyge) was used, and vice versa. White noise from Noisex database [20] was used. For both noise kinds, two different signal to noise output power ratios were selected, 0 dB and 6 dB. A 0 dB output power ratio means that at speakerphones, both signals were replayed with equal powers [4].

All recordings where made at 16000 Hz of sampling frequency with 16 bits quantization. The room used was a sound proof room, with additional plywood reverberation boards in two of the walls to increase reverberation times up to about 260 ms.

To show an example of the algorithm output, the signal aagp 3002 mixed with speech noise at 0 dB was processed with the algorithm. Figure 2 shows the resulting mixed signals as measured by microphones, Fig. 3 shows resulting separated signals and Fig. 4 the source signals (clean, before mixing). This example shows a good separation with large noise reduction, even in a mixture with very strong noise. When listening to the output, the utterance by female speaker can be clearly distinguished, while male noise is heard as a quite low volume background.

In Fig. 5 spectrograms (in dB scale) for mixed signal, separated signal and original signal are shown. A large improvement can be seen, specially in the area where desired signal has low amplitude (towards the end), and also it can be seen how the main structures of the original signal are present in a enhanced way in the separated signal.

To test performance of the algorithm, we have used a speech recognition system to estimate the improvement on word recognition rate before and after separation. For this test, we have used a continuous speech recognizer based on tied Gaussian-mixtures Hidden Markov Models (HMM). This recognizer was trained first with 585 sentences from Minigeo subset of Albayzin database (the training set does not includes any of the phrases used in the test).

After training, we have tested the recognition system with the clean sentences of our test set. To see how the mixing process degrades recognition output we have also evaluated recognition accuracy over the mixtures. We have then applied the separation algorithm without the time-frequency Wiener filter (that test is denoted as J+F because includes only Jade and FastICA separation) and performed recognition accuracy test. Finally, the algorithm for separation including Wiener filter was applied (this test is named J+F+W) and recognition accuracy test was performed on that data. Table 1 shows the results of these tests. In the table, for each test the word recognition accuracy percentage ($W_{ACC}\%$) is

[4] In similar way to standard SNR

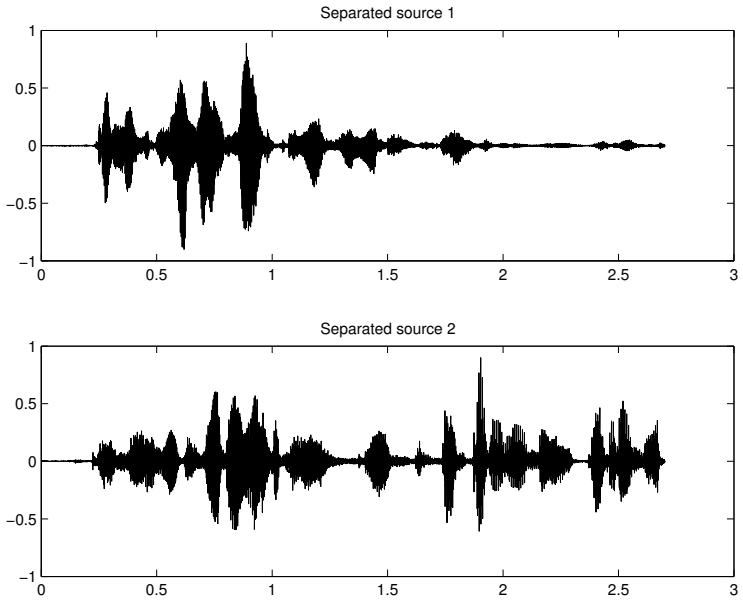

Fig. 3. Separated signals. Top: separated source 1; bottom: separated source 2

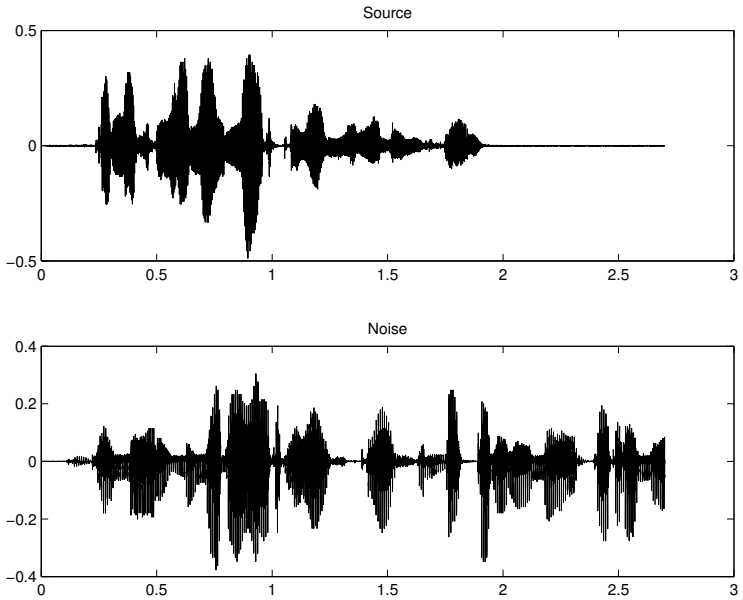

Fig. 4. Source signals. Top: aagp3002, desired source, female; bottom: nyge3110, noise, male.

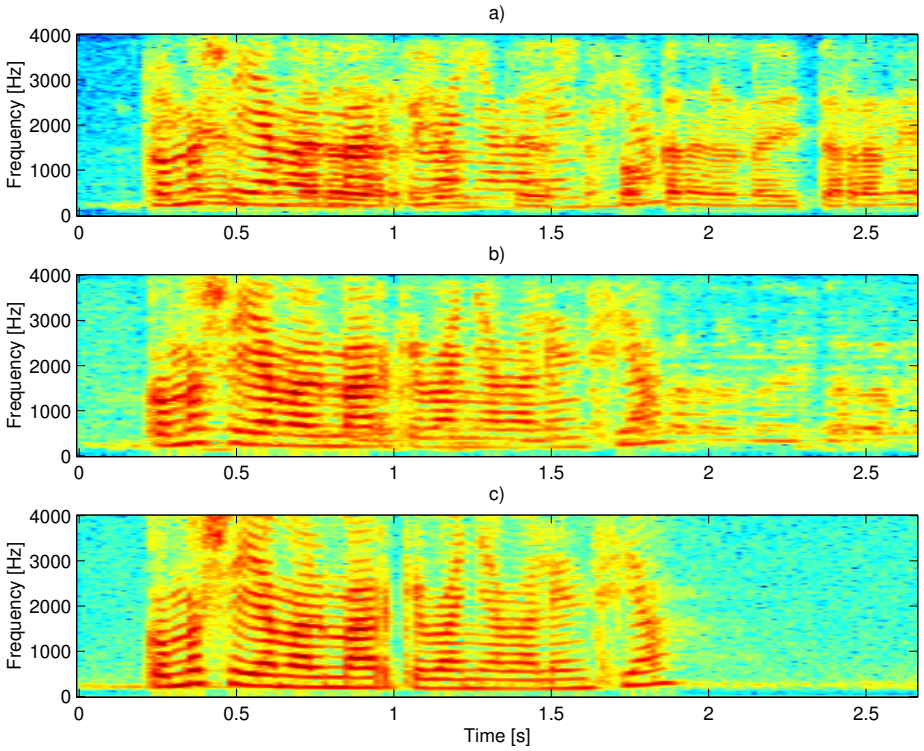

Fig. 5. Spectrograms of a) mixed signal; b) separated signal and c) source signal, for a mixture of speech with speech interference emitted with equal power (power ratio of 0 dB).

calculated as

$$W_{ACC}\% = \frac{N - D - S - I}{N} 100 \tag{14}$$

where N is the number of words in the reference transcription, D is the number of deletion errors (words present in the reference transcription that are not present in the system transcription), S is the number of substitution errors (words that were substituted by others in the system transcription) and I is the number of insertion errors (extra words that were in the system transcription but not in the reference transcriptions). This measure is a more representative figure of recognizer performance than standard word recognition rate [21]. As can be seen from these results, word accuracy rate improvements are in the order of 70%.

4 Conclusions and future works

In this paper an algorithm for blind source separation of convolved sources has been presented. The use of Wiener post-filtering to improve the output of the

Test Environment		$W_{ACC}\%$		
Noise kind	PR dB	Mixtures	J+F	J+F+W
Speech	0	-6.50	38.00	65.50
	6	18.50	68.50	71.50
White	0	3.60	33.00	56.50
	6	12.85	69.50	81.50

Table 1. Word accuracy in robust speech recognition. For reference, with clean sources $W_{ACC}\% = 91.50\%$. PR: Power ratio in dB.

algorithm allows an important reduction in interfering signal power, particularly in the areas were the desired source has low power. As shown by the example in Fig. 2, 3 and 4, the quality can be enhanced to a great extent even in a very bad mixture with equal noise power.

Also, robust speech recognition rates shown a very important improvement in word accuracy, from almost zero percent for mixtures to about 70% after separation with the proposed algorithm. It must be noted that the use of wiener filter has a big effect in word accuracy improvement.

There are some issues that must be addressed for future works. First, we need to explore the capabilities of this algorithm for shorter data. The tests presented here were performed on data with an average duration of 2 seconds. Some applications, like remote controlling of home devices via voice commands or real-time processing for hearing aids, require shorter data to be processed. The algorithms used for separation in each frequency bin need a large amount of data to estimate accurately the statistical properties of signals, so we need to check whether they will still work in cases with less amount of data present.

Finally, some fine tuning of algorithm parameters, like window kind and length used in calculating short time Fourier transform or parameters for the Wiener filter, will be explored.

References

1. Kahrs, M., Brandenburg, K., eds.: Applications of Digital Signal Processing to Audio and Acoustics. The Kluwer International Series In Engineering And Computer Science. Kluwer Academic Publishers (2002)
2. Kinsbury, B., Morgan, N.: Recognizing reverberant speech with RASTA-PLP. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. (1997) 1259–1262
3. Benesty, J., Makino, S., Chen, J., eds.: Speech Enhancement. Signals and Communication Technology. Springer (2005)
4. Cichocki, A., Amari, S.i.: Adaptive Blind Signal and Image Processing. Learning Algorithms and applications. John Wiley & Sons (2002)
5. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Inc. (2001)
6. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Transactions on Speech and Audio Processing **8**(3) (2000) 320–327

7. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing **41**(1-4) (2001) 1–24
8. Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., Saruwatari, H.: Equivalence between Frequency-Domain blind source separation and Frequency-Domain adaptive beamforming for convolutive mixtures. EURASIP Journal on Applied Signal Processing **2003**(11) (2003) 1157–1166
9. Mitianoudis, N., Davies, M.: New fixed-point ica algorithms for convolved mixtures. In: Proceedins of the Third International Conference on Independent Component Analysis and Source Separation. (2001) 633–638
10. Prasad, R., Saruwatari, H., Lee, A., Shikano, K.: A fixed-point ica algorithm for convoluted speech signal separation. In: Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation. (2003) 579–584
11. Gotanda, H., K. Nobu, Koya, T., Kaneda, K., Ishibashi, T., Haratani, N.: Permutation correction and speech extraction based on split spectrum through fastica. In: Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation. (2003) 379–384
12. Douglas, S.C., Sun, X.: Convolutive blind separation of speech mixtures using the natural gradient. Speech Communication **39**(1-2) (2003) 65–78
13. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Transactions on Speech and Audio Processing **12**(5) (2004) 530–538
14. Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. IEEE Transactions on Speech and Audio Processing **11**(2) (2003) 109–116
15. Makino, S., Sawada, H., Mukai, R., Araki, S.: Blind Source Separation of Convolutive Mixtures of Speech in Frequency Domain. IEICE Trans Fundamentals **E88-A**(7) (2005) 1640–1655
16. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non Gaussian signals. IEE Proceedings-F **140** (1993) 362–370
17. Bingham, E., Hyvarinen, A.: A fast fixed-point algorithm for independent component analysis of complex valued signals. International journal of Neural Systems **10**(1) (2000) 1–8
18. Huang, Y.A., Benesty, J., eds.: Audio Signal Processing for next-generation multimedia communication systems. Kluwer Academic Press (2004)
19. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., C. Nadeu: Albayzin speech database design of the phonetic corpus. Technical report, Universitat Politècnica de Catalunya (UPC), Dpto. DTSC (1993)
20. Varga, A., Steeneken, H.: Assessment for automatic speech recognition II NOISEX-92: A database and experiment to study the effect of additive noise on speech recognition systems. Speech Communication **12**(3) (1993) 247–251
21. Yung, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (for HTK Version 3.3). Cambridge University Engineering Department, Cambridge. (2005)