Auditory-based Subband Blind Source Separation using Sample-by-Sample and Infomax Algorithms

Abderraouf Ben Salem[1], Sid-Ahmed Selouani[2], and Habib Hamam[3] *

Abstract—**We present a new subband decomposition method for the separation of convolutive mixtures of speech. This method uses a sample-by-sample algorithm to perform the subband decomposition by mimicking the processing performed by the human ear. The unknown source signals are separated by maximizing the entropy of a transformed set of signal mixtures through the use of a gradient ascent algorithm. Experimental results show the efficiency of the proposed approach in terms of signal-to-interference ratio. Compared with the fullband method that uses the Infomax algorithm, our method shows an important improvement of the output signal-to-noise ratio when the sensor inputs are severely degraded by additive noise.**

Keywords: blind source separation, subband decomposition, ear model, convolutive sources, Infomax algorithm.

1 Introduction

Blind source separation (BSS) is an approach trying to separate mixed sources, assumed to be statistically independent, without any *a priori* knowledge about original source signals $s_j(n), j \in \{1, \cdots, N\}$ but using only observations $x_i(n), i \in \{1, \cdots, M\}$ through M sensors. Such signals are instantaneously or convolutively mixed. In this paper, we are concerned with the convolutive case, i.e. the BSS of convolved sources of speech, where source signals are filtered by impulse responses $h_{ij}(n)$, from source j to sensor i. Mixtures in that case can be expressed under a vector notation as:

$$\mathbf{X}(n) = \sum_{k=0}^{\infty} \mathbf{H}(k)\mathbf{S}(n-k), \tag{1}$$

where $\mathbf{X}(n) = [x_1(n), \cdots, x_M(n)]^T$ is a vector of mixtures, $\mathbf{S}(n) = [s_1(n), \cdots, s_N(n)]^T$ is a vector of speech sources, and $\mathbf{H}(k) = [h_{ij}(k)], (i,j) \in \{1, \cdots, M\} \times \{1, \cdots, N\}$ is a matrix of FIR filters. To blindly estimate the sources, an unmixing process is carried out, and the estimated sources $\mathbf{Y}(n) = [y_1(n), \cdots, y_N(n)]^T$ can be written as:

$$\mathbf{Y}(n) = \sum_{k=0}^{L-1} \mathbf{W}(k)\mathbf{S}(n-k) \tag{2}$$

where $\mathbf{W}(k) = [w_{ij}(k)], (i,j) \in \{1, \cdots, M\} \times \{1, \cdots, N\}$ is the unmixing matrix linking the j-th output $y_j(n)$ with the i-th mixture $x_i(n)$. Such matrix is composed of FIR filters of length L. Each element is defined by the vectors $w_{ij}(k) = [w_{ij}(0), \cdots, w_{ij}(L-1)], \forall (i,j) \in \{1, \cdots, M\} \times \{1, \cdots, N\}$.

To mitigate problems in both time and frequency domains, we present an approach for separating convolutive mixtures based on subband decomposition, referred to as *Subband BSS*. Subband BSS has many advantages compared to the other frequency-Domain BSS approaches regarding the well-known *Permutation ambiguity* of frequency bins [2]. In fact, the subband BSS's permutation problem is quite less critical since the number of subbands that could be permuted is obviously less than the frequency bins. In addition, using a decimation process for each subband can considerably reduce the computational load if compared with time-domain approaches (which could be computationally demanding task in the case of real-room mixtures).

Many subband BSS methods were proposed [1, 6, 7, 8]. In [1], the subband analysis/synthesis system uses a polyphase filterbank with oversampling and single side band modulation. In low frequency bands, longer unmixing filters with overlap-blockshift are used. In [6], the subband analysis filterbank is basically implemented as a cosine-modulated prototype filter. This latter is designed as a truncated sinc(.) function weighted by a Hamming window. In [8], the impulse responses of the synthesis filters are based on the extended lapped transform and are defined by using the cosine modulation function. In the approach reported in [7], analysis filters are obtained by a generalized discrete Fourier transform. Analysis and synthesis filters are derived from a unique prototype filter which can be designed by iterative least-squares

*[1] Canadian University of Dubai, Dubai, UAE. Email: raouf@cud.ac.ae. [2] Université de Moncton, campus de Shippagan E8S 1P6 NB, Canada Email:selouani@umcs.ca. [3] Université de Moncton, campus de Moncton E1A 3E9 NB, Canada Email:habib.hamam@umoncton.ca.

algorithm with a cost function that considers a stopband attenuation.

Throughout this paper, we present a new framework for the BSS of convolutive mixtures based on subband decomposition using an ear-model based filterbank and information maximization algorithm.

2 Proposed method

In this section, we define the subband decomposition using the modeling of the mid-external ear and the basilar membrane that aims at mimicking the human auditory system (HAS). Afterwards, the learning rule wich performs the sources' separation will be introduced.

2.1 Subband decomposition

The proposed modeling of HAS consists of three parts that simulate the behavior of the mid-external ear, the inner ear and the hair-cells and fibers. The external and middle ear are modeled using a bandpass filter that can be adjusted to signal energy to take into account the various adaptive motions of ossicles. The model of inner ear simulates the behavior of the basilar membrane (BM) that acts substantially as a non-linear filter bank. Due to the variability of its stiffness, different places along the BM are sensitive to sounds with different spectral content. In particular, the BM is stiff and thin at the base, but less rigid and more sensitive to low frequency signals at the apex. Each location along the BM has a characteristic frequency, at which it vibrates maximally for a given input sound. This behavior is simulated in the model by a cascade filterbank. The number of filterbank depends on the sampling rate of the signals and on other parameters of the model such as the overlapping factor of the bands of the filters, or the quality factor of the resonant part of the filters. The final part of the model deals with the electro-mechanical transduction of hair-cells and afferent fibers and the encoding at the level of the synaptic endings [9].

2.1.1 Mid-External Ear

The mid-external ear is modeled using a bandpass filter. For a mixture input $x_i(k)$, the recurrent formula of this filter is given by:

$$x_i^{'}(k) = x_i(k) - x_i(k-1) + \alpha_1 x_i^{'}(k-1) - \alpha_2 x_i^{'}(k-2), \quad (3)$$

where $x_i^{'}(k)$ is the filtered output, $k = 1, \cdots, K$ is the time index and K is the number of samples in a given block. The coefficients α_1 and α_2 depend on the sampling frequency F_s, the central frequency of the filter and its Q-factor.

2.1.2 Mathematical Model of the Basilar Membrane

After each frame is transformed by the mid-external filter, it is passed to the cochlear filter banks whose frequency responses simulate those of the BM for an auditory stimulus in the outer ear. The formula of the model is as follows:

$$x_i^{''}(k) = \beta_{1,i} x_i^{''}(k-1) - \beta_{2,i} x_i^{''}(k-2) + G_i[x_i^{'}(k) - x_i^{'}(k-2)], \quad (4)$$

and its transfer function can be written as:

$$H_i(z) = \frac{G_i\,(1 - z^{-2})}{1 - \beta_{1,i}\, z^{-1} + \beta_{2,i}\, z^{-2}}, \quad (5)$$

where $x_i^{''}(k)$ is the BM displacement which represents the vibration magnitude at position δ_i and constitutes the BM response to a mid-external sound stimulus $x_i^{'}(k)$. The parameters G_i, $\beta_{1,i}$ and $\beta_{2,i}$, respectively the gain and coefficients of filter or channel i, are functions of the position δ_i along the BM. N_c cochlear filters are used to realize the model. These filters are characterized by the overlapping of their bands and a large bandwidth. We assume that the BM has a length of 35 millimeters which is approximately the case for humans. Thus, each channel represents the state of an approximately $\Delta = 1.46$ mm of the BM. The sample-by-sample algorithm providing the outputs of the BM filters is given as follows.

```
**Initialize** $f_x = (F_s\,\Delta x)^2; H_0 = 0; r_{i,j} = 0; E_0 = 0.$
**For** $i = 1$ **to** $N_c$ **do**
  $x_i = i\,\Delta x; v = e^{-106.5\,x_i}; F_i = 7100\,v - 100;$
  $C_i = \frac{(27\,v)^2}{f_x}; Q_i = (-8300\,x_i + 176.3)\,x_i + 4;$
  $G_i = e^{-80\,x_i}; u = e^{-\frac{\pi\,F_i}{F_s\,Q_i}}; \beta_{1,i} = 2\,u\,cos(\frac{2\,\pi\,F_i}{F_s});$
  $\beta_{2,i} = u^2; E_i = \frac{1}{1+(2-E_{i-1})\,C_i}; A_i = E_i\,C_i.$
**EndDo**
**For** $k = 1$ **to** $K$ **Do**
   **For** $i = 1$ **to** $N_c$ **Do**
     $H_i = [G_i\,(s^{'}(k) - s^{'}(k-2)) + \beta_{i,2}\,r_{1,i} -$
          $\beta_{2,i}\,r_{i,1}]\,E_i + H_{i-1}\,A_i$
   **EndDo**
   **For** $i = 1$ **to** $N_c$ **Do**
     $r_{1,i} = A_i\,r_{i+1,3} + H_i; y^{'}_i(k) = r_{i,3}$
   **EndDo**
   **For** $i = 1$ **to** $N_c$ **Do**
     **For**$j = 1$ **to** $2$ **Do**
       $r_{i,j} = r_{i,j+1}$
     **EndDo**
   **EndDo**
**EndDo**
```

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

2.2 Learning algorithm

After performing the subband decomposition, the separation of convolved sources per subband is done by the Infomax algorithm. Infomax was developed by Bell and Sejnowski for the separation of instantaneous mixtures [3]. Its principle consists of maximizing output entropy or minimizing the mutual information between components of $\mathbf{Y}$. It is implemented by maximizing, with respect to $\mathbf{W}$, the entropy of $\mathbf{Z} = \Phi(\mathbf{Y}) = \Phi(\mathbf{WX})$. Thus, the Infomax contrast function is defined as:

$$C(\mathbf{W}) = H(\Phi(\mathbf{WX})), \qquad (6)$$

where $H(.)$ is the differential entropy, which can be expressed as $H(a) = -E[Ln(f_a(a))]$, where $f_a(a)$ denotes the probability density function of a variable a. The generalization of Infomax for the convolutive case by using a feedforward architecture is introduced in the proposed method. Both causal and non-causal FIR filters are performed in our experiments. With real-valued data for vector $\mathbf{X}$, entropy maximization algorithm leads to the adaptation of unmixing filter coefficients with a stochastic gradient ascent rule using a learning steepest μ. Then, the weights are defined as follows.

$$\mathbf{W}(0) = \mathbf{W}(0) + \mu([\mathbf{W}(0)]^{-T} - \Phi(\mathbf{Y}(n))\mathbf{X}^T(n)), \qquad (7)$$

and,

$$w_{ij}(k) = w_{ij}(k) - \mu\, \Phi(y_i(n))x_j(n-k); \forall k \neq 0, \qquad (8)$$

where $\mathbf{W}(0)$ is a matrix composed of unmixing FIR filters coefficients as defined in Section 1, $\mathbf{Y}(n)$ and $\mathbf{X}(n)$ are the separated sources and the observed mixtures, respectively. $\Phi(.)$ is the score function of y_i which is a non-linear function approximating the cumulative density function of sources, as defined in Eq. 9, where $p(y_i)$ denotes the probability density function of y_i.

$$\Phi(y_i(n)) = \frac{\frac{\delta p(y_i(n))}{\delta y_i(n)}}{p(y_i(n))}. \qquad (9)$$

The block diagram of the proposed method is given in Fig. 1. The input signals, that are the set of mixtures, are firstly processed by the mid-external ear introduced by Eq. 3, then outputs are passed through a filterbank representing the cochlear part of the ear. A decimation process is then performed for each subband output. Such decimation is useful for many reasons. First, it improves the convergence speed because input signals are more whitened than the time domain approach. Second, the wanted unmixing filter length will be reduced by a factor of $\frac{1}{M}$, where M is the decimation factor. After performing decimation, we group a set of mixtures belonging to the same cochlear filter to be the input of the unmixing stage. This latter gives separated sources of each subband that are upsampled by a M factor. The same filterbank is used for the synthesis stage. The estimated sources are added from different synthesis stages.

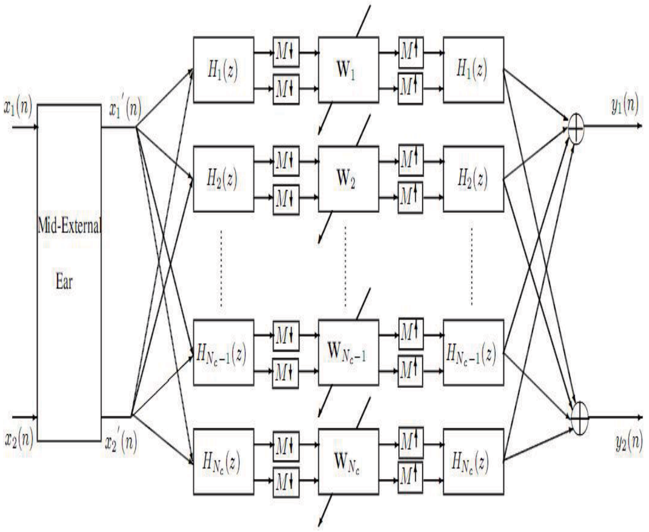

Figure 1: The ear-based framework for the subband BSS of convolutive mixtures of speech.

3 Experimental results

In order to evaluate the separation performance given by the proposed ear-based subband BSS method, a set of experiments have been carried out. In our experiments, we used as sources speech data containing two sentences spoken by a male and female speakers, those signals are at sampling rate of 8 kHz, each one is an excerpt of 6 seconds. Infomax algorithm has several parameters: the mixture signals are segmented into blocks; each block is a size of 35 samples, and the learning rate was fixed to $\mu = 0.001$. Further, $\Phi(u) = \frac{1}{1+e^{-u}}$ was used to approximate the cumulative density function. Such sources are convoluted with impulse responses modeling the Head Related Transfer Function (HRTF) [5]. We tested our overall framework with a mixing filters measured at the ears of a dummy head, illustrated by figure 2 . We selected impulse responses associated with source positions defined by 30- and 80-degree angles in relation to the dummy head.

To evaluate the performance, the Signal to Interference Ratio (SIR) is used [4]. This reliable measurement is defined by:

$$SIR = 10\, log_{10} \frac{||s_{target}||^2}{||e_{interf}||^2}, \qquad (10)$$

where $s_{target}(n)$ is an allowed deformation of the target source $s_i(n)$, $e_{interf}(n)$ is an allowed deformation of the

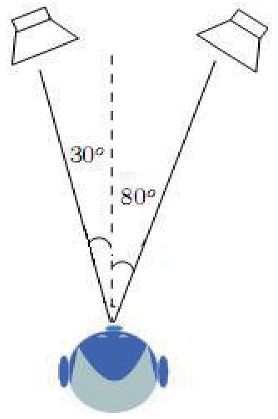

Figure 2: The convolutive model with source positions at 30-and 80-degree angles in relation to the dummy head.

sources which accounts for the interference of the unwanted sources. Those signals are derived from a decomposition of a given estimated source $y_i(n)$ of a source $s_i(n)$.

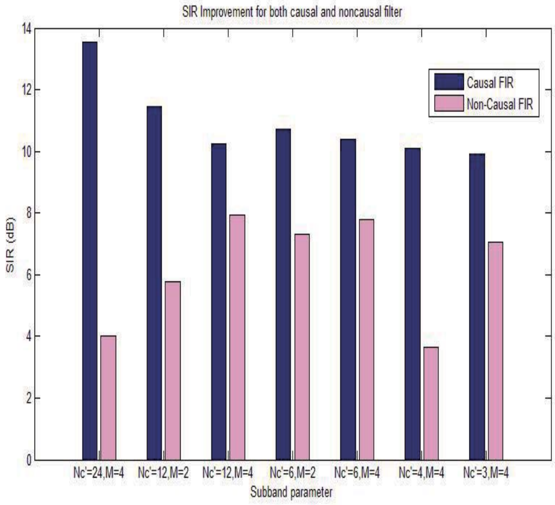

Figure 3: SIR improvement for both causal and noncausal filters. We denote by $N_c^{'}$ the number of filters that have been used among N_c filters and M is decimation factor.

Different configurations of the subband analysis and synthesis stages as well as of the decimation factor have been tested. The number of subbands was fixed at 24. Through our experiments we observed that when we keep the whole number of subbands, the results were not satisfactory. In fact, we noticed that some subbands in high frequencies are not used, and therefore this causes distortions on the listened signals. However, as shown in Figure 3, the best performance was achieved for $N_c^{'} = 24$ and $M = 4$. In addition to the use of causal FIR filters, we adapted unmixing stage weights for non-causal FIR by centering the L taps. From Figure 3, we observe that causal FIR yields to good results in SIR improvement when compared with non-causal one. Another set of experiments have been carried out to evaluate the performance in the presence of an additive noise in sensors. We used the Signal-to-Noise-Ratio (SNR) which is defined in [4], by:

$$SNR = 10\, log_{10} \frac{||s_{target} + e_{interf}||^2}{||e_{noise}||^2}, \qquad (11)$$

where e_{noise} is an allowed deformation of the perturbating noise, s_{target} and e_{interf} were defined previously. Figure 4 shows the SNR improvement using our subband decomposition, comparing to the fullband method, i.e. Infomax algorithm in convolutive case.

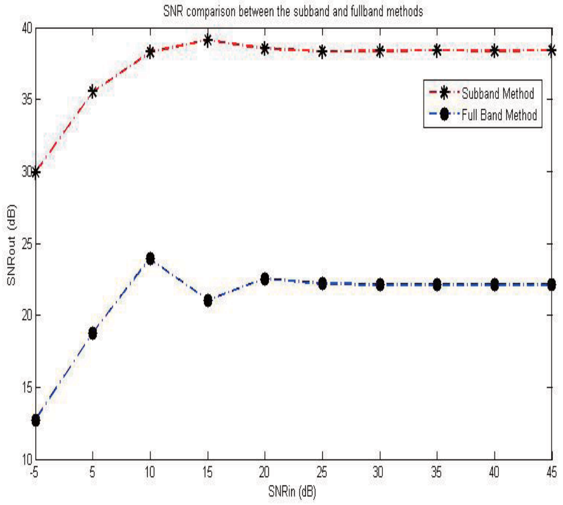

Figure 4: SNR comparison between the subband and fullband methods.

4 Conclusion

An ear-based subband BSS approach was proposed for the separation of convolutive mixtures of speech. The results showed that using a subband decomposition that mimics the human perception and using the Infomax algorithm yields to better results than the fullband method. Experimental results show the high efficiency of the new method in improving the SNR of unmixed signals in the case of noisy sensors. It is worth noting that an important advantage of the proposed technique is that it uses a simple time-domain sample-by-sample algorithm to perform the decomposition.

References

[1] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Subband-Based Blind Separation for Convolutive Mixtures of Speech", *IEICE Trans. Fundamentals*, Vol. E88-A, No. 12, pp. 3593-3603, December 2005.

[2] S. Araki, S. Makino, T. Nishikawa, and H. Saruwarati, "Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixture of Speech", *IEEE-ICASSP conference*, pp. 2737-2740, 2001.

[3] A. J. Bell and T. J. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution", *Neural Computation*, pp. 1129-1159, 1995.

[4] C. Fevotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide", IRISA, Rennes, France, *Technical Report 1706*, 2005. [Online]. Available: *http://www.irisa.fr/metiss/bss_eval.*

[5] B. Gardner and K. Martin, "Head Related Transfer Functions of a Dummy Head", [Online]. Available: *http://sound.media.mit.edu/ica-bench/.*

[6] K. Kokkinakis and P. C. Loizou, "Subband-Based Blind Signal Processing for Source Separation in Convolutive Mixtures of Speech", *IEEE-ICASSP Conference*, pp. 917-920, 2007.

[7] H.-M. Park, C. S. Dhir, S.-H. Oh, and S.-Y. Lee, "A Filter Bank Approach to Independant Component Analysis for Convolved Mixtures", *Neurocomputing* , pp. 2065-2077, 2006.

[8] I. Russel, J. Xi, A. Mertins, and J. Chicharo, "Blind Source Separation of Non-Stationary Convolutively Mixed Signals in the Subband Domain", *IEEE-ICASSP Conference*, pp. 481-484, 2004.

[9] H. Tolba, S.A Selouani, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multistream paradigm ", *IEEE-ICASSP Conference 2002*, pp. 837-840, 2002.