CONCURRENT SPEAKERS SEPARATION THROUGH BINAURAL PROCESSING OF STEREO RECORDINGS

Joaquin Gonzalez-Rodriguez, Santiago Cruz-Llanas and Javier Ortega-Garcia
DIAC-EUIT.Telecomunicacion-Universidad Politecnica de Madrid
Campus Sur - 28031 Madrid SPAIN
jgonzalz@diac.upm.es
http://www.atvs.diac.upm.es

ABSTRACT

In this contribution, we address the problem of speech separation from concurrent speakers through the use of a binaural processor. In this system, we take advantage of both ITD's (Inter-aural Time Differences) and ILD's (Inter-aural Level Difs.) present in the binaural input signals to perform spatial localisation and discrimination. However, in applications as Forensic Acoustics where hidden recordings are desired we cannot make use of head simulators, but we can also obtain ITD's and ILD's through stereo recordings with a cardioid microphone pair. In this way, after a new training of the processor, we are able to extract the information arriving from the desired direction from a simple stereo recording. Several experiments with static and moving speakers have been performed, and some audio results on concurrent speakers separation are also included.

Keywords: Binaural Processing, Speech Enhancement, Auditory Modeling, Spatial Localization, Forensic Acoustics.

1. INTRODUCTION

Speech separation from concurrent speakers is an application of special interest to the Forensic Acoustics community, where hidden recordings usually include speech from other persons different from the target speaker. Even though speech enhancement systems have been designed to cope both with additive and convolutional noise, when the disturbing signal is also a speech signal from a concurrent speaker, conventional speech enhancement methods are not able to extract the desired signal from the mixed input. In our work, we take advantage of the spatial characteristics of the sound field together with some properties of the receivers to extract the information from one static or moving speaker.

Previous systems based on binaural processors [1] depend on a binaural input signal, obtained through actual recordings with a head simulator, or simulating the recording process through convolution with the Head Related Transfer Functions (HRTFs) for the correspondent angles. However, the binaural processor just need to know some ITD's and ILD's but not necessarily the same as a human listener, in order to perform spatial localisation and spectral discrimination.

In this work, ITD's and ILD's are obtained with a simple pair of opposite cardioid microphones, with a separation of 20 centimetres. As a consequence, the processor has to be retrained, but it is able from now on to perform localisation and discrimination from simple stereo recordings.

The paper is organised as follows. First, we will describe in section 2 the binaural processor that we have implemented, discussing in section 3 the advantages of the cardioid-pair input signal instead of the conventional binaural one. Section 4 will describe the structure of the speech enhancement system based in the binaural processor, and finally in section 5 we will present some experiments on concurrent speakers separation.

2. SYSTEM DESCRIPTION

The binaural processor we have implemented in this work is based in the system described in [1], where the different modules try to imitate the human auditory abilities that finally allow us to discriminate between different signal sources and background noises. These modules are described in the following subheadings.

2.1. Inner ear simulation

The input to the system is a binaural signal that includes the ITD's and ILD's that will allow the system to perform properly. This signal can be obtained from actual recordings with a head simulator or simulating the acoustic propagation in the room and the influence of the human head through convolution with the room impulse response [4] and the HRTF's (Head Related

Transfer Functions). In each one of both channels the signal is processed through a Zwicker critical-band filter-bank where the firing nerve probabilities are obtained trough half-wave rectification of the filterbank outputs.

In figure 1, we can observe the energy frequency contour for two isolated different sounds with no acoustic propagation nor head influence.

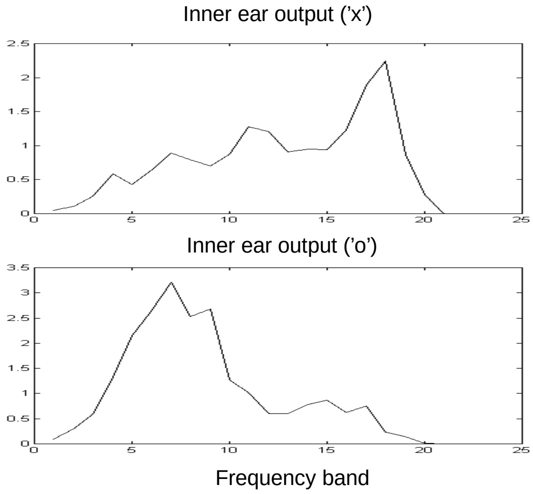

Figure 1.- *Average energy in the different frequency bands for 'x' and 'o' in spanish.*

2.2. Crosscorrelation module

The ITD's observed between the two inner ear outputs in each frequency band are the key to the crosscorrelation module. However, with a simple crosscorrelation it is not possible to obtain valid estimations of the spectral content arriving in each different direction, as we can see in figure 2(a). But if this operation is performed through contralateral inhibition [2] of the neural excitation patterns, taking also advantage of the ILD's, we can obtain a spatial-frequency representation of the sound field arriving to the system in each analysis frame, as in figure 2(b).

2.3. Localisation module

Combining the information from the different nerve excitation patterns, the localisation module extracts an estimation of the arrival angle of the different sources to the binaural processor (fig. 3). This module is trained with inputs from different known angles, where the estimations from the different bands are weighted in order to give the correct estimation. Of course, in the case of a different receiver, as the case of the cardioid-microphone pair that we will see later in this paper, this module has to be retrained for the new acoustic situation.

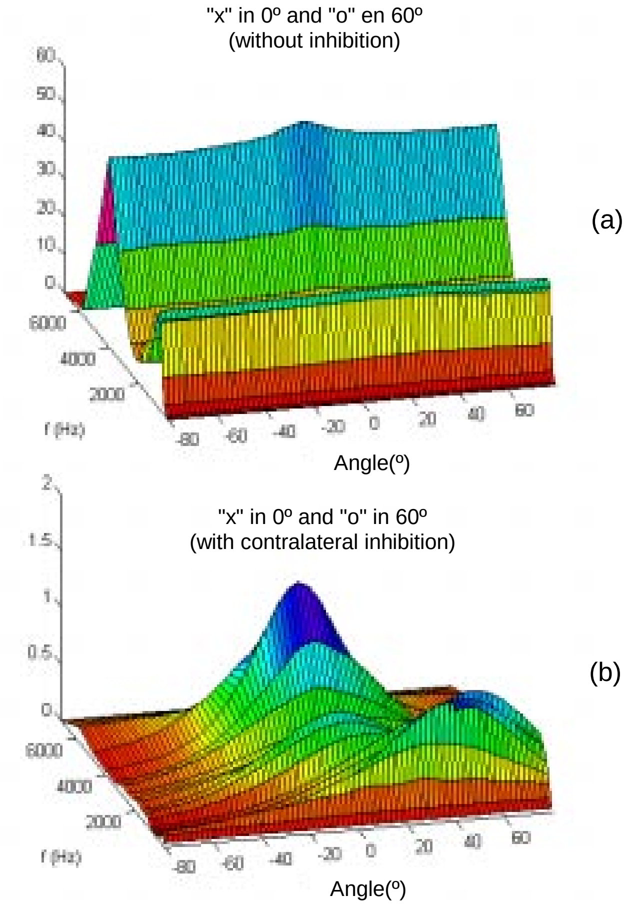

Figure 2.- *Neural excitation patterns from the correlation module (a) without and (b) with contralateral inhibition, for two simultaneous sounds from different directions*

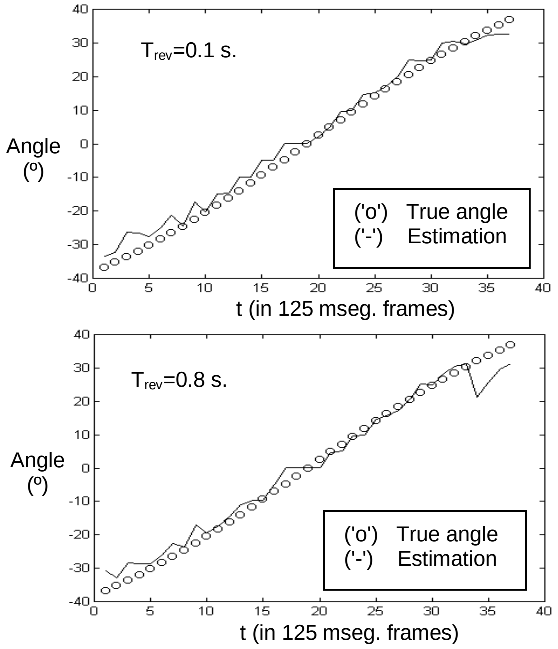

Figure 3.- *Localisation experiments for a moving speaker in front of the receiver with different reverberation times.*

2.4. Discrimination module

This module extracts the desired spectral information combining the neural excitation patterns obtained in the crosscorrelation module and the estimation of the arrival angle from the localisation module. We can observe the extremely efficient performance of the system comparing the results presented in figure 1, where both signals are isolately presented to the inner ear module, and those presented in figure 4, where the binaural input includes two concurrent signals, present in both input channels, but including ITD's and ILD's that allow the system to perform the separation.

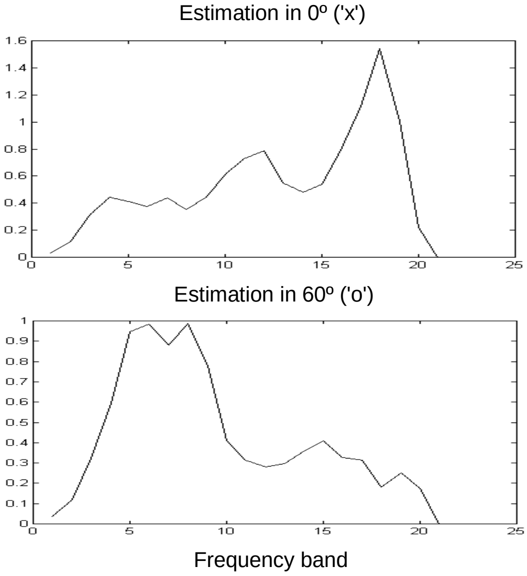

Figure 4.- *Estimation of the frequency content of two concurrent sounds with different arrival angles*

3. BINAURAL VERSUS STEREO RECORDINGS

The use of a head simulator is not mandatory to obtain ITD's and ILD's at the input to the binaural system.

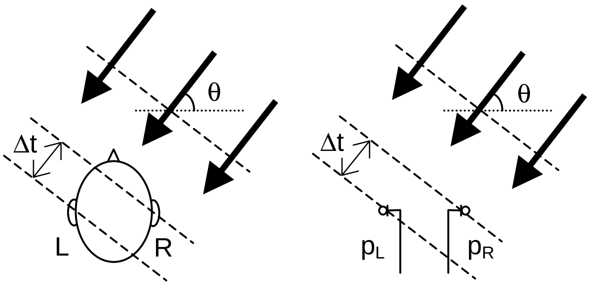

Figure 5.- *Obtention of the ITD's and ILD's with the binaural and the cardioid-pair receivers*

Instead, we propose the use of a pair of cardioid microphones, aimed at opposite directions and with a 20 centimetres separation between them. Of course, the localisation module has to be retrained, but we obtain with a simple pair of microphones different ITD's and ILD's for each one of the possible incoming directions (we have to note that the binaural processor just need ITD's and ILD's to work properly, but not necessarily the same ITD's and ILD's as a human to perceive it 'naturally').

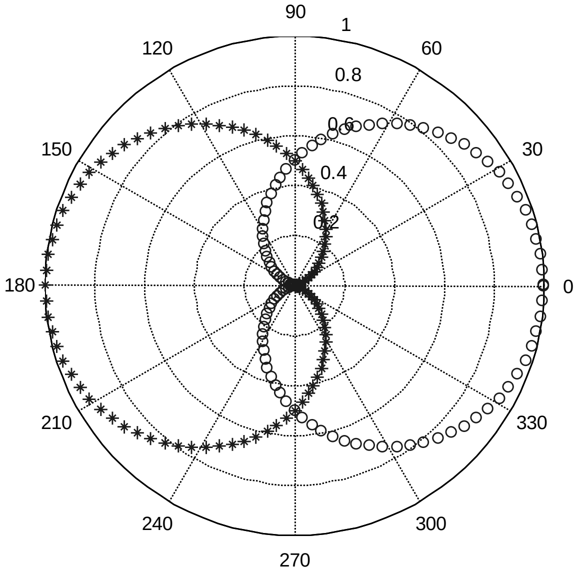

Figure 6.- *Directivity patterns of the cardioid microphone pair*

4. SPEECH ENHANCEMENT THROUGH BINAURAL PROCESSING

The speech separation will be performed through Wiener filtering of one of the input signals. Then, in order to design the filter, we need a spectral estimator of the desired components and an estimator of the power spectral density at the input. The former estimator is obtained, frame by frame, from the spatial-frequency distribution of the neural excitation patterns, where the frequency content of each speaker, fully mixed in the two input signals, is now placed in different spatial positions corresponding to the different input angles of each speaker, which allow the system to extract the information correspondent just to the target speaker. Conversely, the latter estimator is directly obtained from the input signal. Moreover, any coherent noise component arriving to the microphones [3] will not be included in the first estimator, and then will be suppressed in the filtered output signal.

Once we have available the filter, the input sequence is frequency-domain filtered in each analysis frame, and again converted to the time domain through inverse frequency transformation and overlap-add processing, as shown in figure 7.

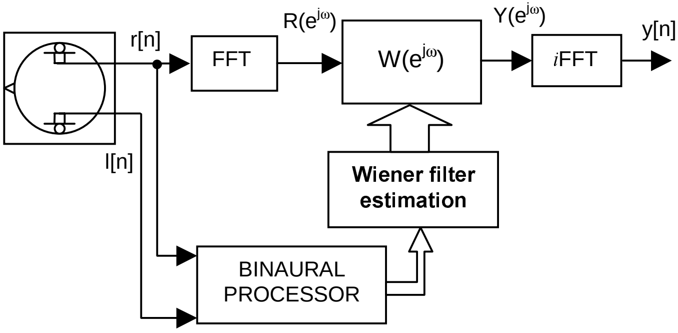

Figure 7.- *Obtention of the ITD's and ILD's with the binaural and the cardioid-pair receivers*

5. EXPERIMENTS AND RESULTS

Different experiments have been performed both with concurrent speakers and/or noise. The acoustic propagation from the desired and undesired sources have been simulated through convolution with the impulse responses obtained with the image method [4], obtaining then the two mixed input signals to the system, taking into account the HRTF's or the directivity patterns of the cardioid microphones in each case..

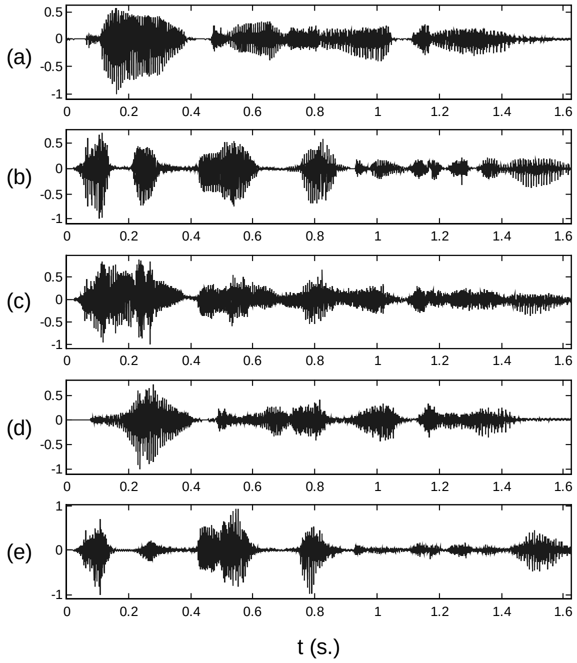

Figure 8.- *(a) Spk A, (b) Spk B, (c) Left channel from the binaural input, (d) Estimation of Spk A, (e) Estimation of Spk B*

The results obtained when just one speaker and noise are present are equivalent to those obtained with other speech enhancement systems, as spectral subtraction [5]. However, when two speakers are present at the same time, the separation obtained is excellent with very little spectral distortion and perfect intelligibility [6].

In figure 8 we can see an example of the performance of the system from a binaural input (the left channel of it is shown in fig. 6c), where two speakers (A and B) are concurrently speaking to the system from different positions.

6. CONCLUSION

In this contribution, a new approach to speaker separation from just two microphones has been described. Based on the spatial-frequency decomposition obtained through the binaural processor, the Wiener filtered output signal is free from the undesired speaker and any directional noise arriving to the microphone pair with no noticeable distortion.

ACKNOWLEDGEMENTS

The authors wish to thank Ignacio Lopez-Monclus for his excellent work on the development of the binaural processor in his MSc Thesis.

7. REFERENCES

[1] M. Bodden and T.R. Anderson, "A Binaural Selectivity Model for Speech Recognition", Proc. Eurospeech'95, Madrid, pp 127-130, 1995.

[2] W. Lindemann, "Extension of a Binaural Cross-Correlation Model by Contralateral Inhibition", Journal of the Acoust. Soc. Am. (JASA), vol. 80, pp. 1608-1630, 1986.

[3] J. Gonzalez-Rodriguez, S. Cruz-Llanas and J. Ortega-Garcia., "Coherence-based Subband Decompositon for Robust Speech and Speaker Recognition in Noisy and Reverberant Rooms", Proc. of ICSLP'98, pp. 385-388, Sydney, 1998.

[4] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", JASA, vol. 65 (4), pp. 943-950, 1979.

[5] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Providing Single and Multi-channel Acoustical Robustness to Speaker Identification Systems", Proc. ICASSP'97, pp. 1107-1110, Munich, 1997.

[6] J. Gonzalez-Rodriguez, *Influence and Compensation of the Acoustic Environment in Automatic Speaker Recognition Systems* (in spanish), PhD Thesis Dissertation, Universidad Politecnica de Madrid, March 1999.