Double-vowel Segregation based on a Cochleotopic/AMtopic Map using a Biological Neural Network

Abstract

We propose an auditory scene analyzer for double vowel segregation. Our technique is based on a two-dimensional representation of the signal that generates a cochleotoptic/AMtopic map, which is used to mimic the periodicity analysis performed in the principal monaural pathway at the brain-stem level. In our scheme, the incoming signal is first processed by a cochlear filter bank and the AM modulation envelopes are extracted for the outputs of the bank. The FFT of the output of each channel is then computed. The two-dimensional representation we obtain this way is then applied to a network of bio-inspired neurons. We use three versions of the network: relaxation oscillator, chaotic, and integrate-and-fire networks. The biological neural map acts as a coherence detector, i.e., regions with different oscillation phases are formed based on the onset times of events belonging to different sources. These synchronized regions are used as masks to segregate sources.

Introduction

The Computational Auditory Scene Analysis (CASA) deals with the problem of separating sound sources using psychological and physiological cues. This is in contrast with techniques that are based on mathematical and statistical methods, like "The Beamforming" or ICA (Independent Component Analysis). CASA is done either by expert systems [1][2] or neural networks [3][4].

The basics are laid on the segregation/streaming of sound objects (Gestalt Theory) [5]. Segregation consists of finding elementary objects in the scene and streaming is the " binding" of these objects in order to find which elements belong to which sound source [6].

Our technique uses a pre-processing to extract a two-dimensional representation of the sound, known as Cochleotopic/AMtopic map [9]. This representation is then applied to the network architecture explained below.

Pre-Processing stage

The sound source is processed by a cochlear filter bank followed by an envelope detection. This processing is partly equivalent to the behavior of the basilar membrane and the hair cells in the ear. We used a bank of 24 filters for which the center frequencies range from 300Hz to 4.5kHz. The envelopes of these outputs are extracted afterwards.

The Fast Fourier Transform (FFT) of the envelopes, which partially mimics the behavior of the auditory path at the brain-stem level, are then computed,. Fig.1 shows the output of this whole pre-processing for a mixture of French and

As it can be seen in Fig.1, there is a rough harmonicity in the patterns, which shows the nature of the underlying sound and a criterion to separate audio sources.

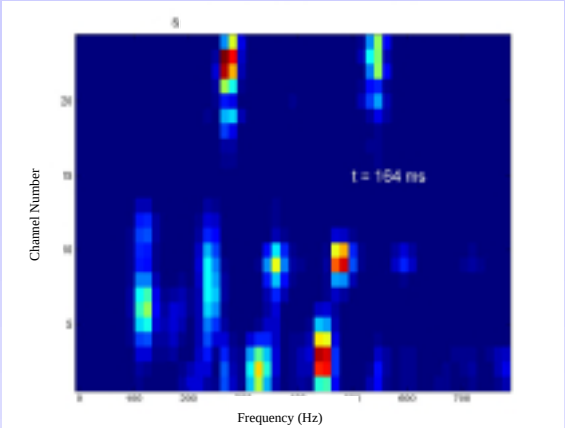

Fig.1 Envelope Spectrogram for a mixture of and

The proposed neural network, extracts this harmonicity in its first layer. A second layer of the network associates each of the channels to one or to the other source based on the harmonicity criterion (geometrical distance between rays in figure 1) .

Architecture of the neural network

The network consists of two layers. The first layer is a map of locally connected Relaxation Neurons [7]. The Envelope spectrogram is applied to this layer. This layer performs the segregation of the scene [5]. In other words, it enhances the harmonicity pattern seen in Fig.1. The second layer is an array of 24 integrate-and-fire neurons [8] (one for each channel). (Fig. 2)

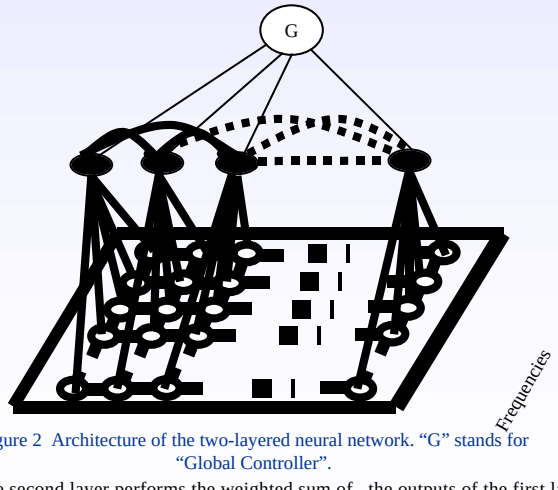

Figure 2 Architecture of the two-layered neural network. "G" stands for "Global Controller".

The second layer performs the weighted sum of the outputs of the first layer neurons along the frequency axis. Due to the harmonicity pattern found in the Fig.1 this will form two different synchronized regions.

Neurons are synchronized among each group and desynchronized in different groups. In this way, the"binding" of the objects extracted in the first layer is done [6] (Fig. 3). We add up the channels for which the second layer corresponding neurons are synchronized, in order to synthesize the initial (non-mixed) speech.

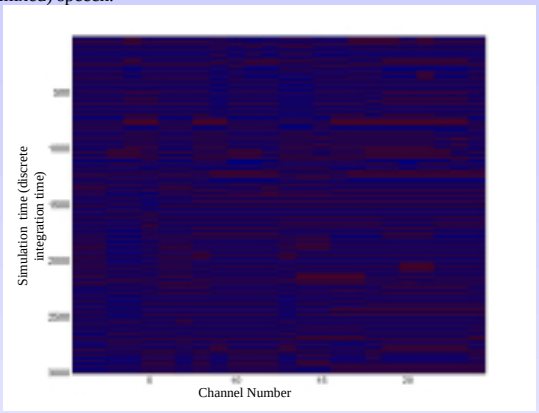

Fig.3 The output of the second layer for only a segment of the vowel

Further Works

More quantitative work has to be done to compare this technique to other available techniques in the literature. This method is mostly tested for vowels (voiced sound). The unvoiced sound case must be studied and compared. The low-frequency channels have a different behavior. The architecture of the network should be enhanced to comply with this difference. This method can be applied also to music mixtures.

Acknowledgement

We would like to thank Romain Balleraud for helping us generate and analyze the Cochleotopic/AMtopic images.

References

[1] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. Speech Communication, 2001, vol. 35, no. 3-4, pp. 141-177.

[2] G. Brown and M. Cooke. Computational auditory scene analysis. Computer Speech and Language, pages 297–336, 1994.

[3] D. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Transactions on Neural Networks, 10(3) :684–697, May 1999.

[4] J. Rouat and R. Pichevar. Nonlinear speech processing techniques for source segregation. In EUSIPCO2002, 2002.

[5] Al Bregman. Auditory Scene Analysis. MIT Press, 1994.

[6] C. Von der Marlsburg and W. Schneider. A neural cocktail-party processor. Biol. Cybernetics, pages 29–40, 1986.

[7] D. L. Wang and D. Terman. Image segmentation based on oscillatory correlataion. Neural Computation, pages II 521– II 525, 1995.

[8] W. Gerstner. Spiking neurons. In W. Maass and C.M. Bishop, editors, Pulsed NeuralNetworks, chapter 1, pages 3–53. MIT Press, 1999.

[9] F. Berthommier and G. Meyer, "Improving of Amplitude Modulation Maps for F0-Dependent Segregation of Harmonic Sounds", In Eurospeech 97