

BINDING OF AUDIO ELEMENTS IN THE SOUND SOURCE SEGREGATION PROBLEM VIA A TWO-LAYERED BIO-INSPIRED NEURAL NETWORK

Ramin Pichevar

GEGI, Univ. de Sherbrooke, QC, Canada
DSA-UQAC, Chicoutimi, QC, Canada
email: pichevar@hermes.usherb.ca

Jean Rouat

GEGI, Univ. de Sherbrooke, QC, Canada
DSA-UQAC, Chicoutimi, QC, Canada
email: jean.rouat@usherbrooke.ca

ABSTRACT

We use a two-layered bio-inspired neural network to segregate sound sources, i.e. double-vowels or intruding noises in speech. The architecture of the network consists of spiking neurons. The spiking neurons in both layers are modeled by relaxation oscillators. The first layer of the network is locally connected, while the second layer is a fully connected network. Our auditory image is based on the reassigned spectrum technique. No prior estimation or knowledge of pitch is necessary for the segregation.

Keywords: CASA(Computational Auditory Scene Analysis), bio-inspired neural networks, cocktail-party effect, sound segregation, Cochleotopic/AMtopic maps.

I. INTRODUCTION

Our proposed technique is a bio-inspired solution to the CASA problem. In this section we will try to clarify some of the key concepts.

A. The "Cocktail-party effect" and CASA

Humans are able to segregate a desired source in a mixture of sounds ("cocktail-party effect"). Experiments have shown that although binaural audition may help improve the segregation performance, human beings are capable of doing the segregation even with one ear or when all the sources come from the same spatial location(for instance, when someone listens to a radio broadcast) [1]. This is in contrast with the statistical (i.e., the Independent Component Analysis (ICA)) or signal processing techniques (spatial localization) that are based on the use of two or more microphones. These differences between mathematical methods and psychoacoustic findings suggest that the human brain and the auditory pathway process the information differently. Using the knowledge acquired in visual scene analysis and by making an analogy between the vision and the audition, Bregman developed the key notions of the "Auditory Scene Analysis" [1]. Two of the most

important aspects in ASA are the "segregation" and "streaming" of sound sources. The segregation step partitions the auditory scene into some fundamental auditory elements and the streaming is the binding of these elements in order to reproduce the initial sound sources. These two stages are influenced by top-down processing (or schema-driven). The aim in Computational Auditory Scene Analysis (CASA) is to develop computerized methods for solving the sound segregation problem by using psychoacoustical and physiological cues. Previous works that deal with the CASA problem are based either on expert systems (for a review see [2]) or on neural networks [3], [4].

B. Binding of Auditory Sources

Different features of speech are extracted in different parts of the brain. These different features are "bound" together in the brain. We can assume that the sound segregation is a generalized classification problem, in which we want to bind features extracted in different sections of our neural network map. This generalized classification problem was first addressed by Rosenblatt [5]. Suppose that there are two sources and two features to classify. In the case of static neurons (i.e., perceptrons) we can implement a network in which one neuron is triggered by the existence of S1(Source 1), another one by S2 (Source 2), a third one with F1 (a specific speech feature like a modulation frequency, an onset/offset time, a modulation phase, etc.) and the final one with F2 (another feature). Suppose that we apply the preprocessed signal from S1 (which contains the F1 feature) to the network. Neurons "S1" and "F1" will be activated. Now suppose that a mixture of S1 and S2 (which contains both the F2 and F1 features) are applied to the network. In this case, all four neurons will be turned on. The network is now confused: It doesn't know whether S1 goes with F1 or with F2. In other words, which one is the correct binding [(S1, F1), (S2, F2)] or [(S1, F2), (S2, F1)]? Three solutions to this problem are proposed in the literature:

- The most straightforward solution to this problem is the hierarchical coding of the information. One neuron is triggered when the stimulus (S1, F1) is present and another one turns on when the input (S1, F2) is applied and so on for all the possible combinations [6]. The problem with this approach is an exponential increase in the number of neurons with a rise in the number of classes and a lack of autonomy for new (not previously seen) classes. Its advantage is its fast response time (classification time).
- Another solution to the aforementioned problem is the use of attentional models [7]. In this method attention is focused on one of the elements in the stimulus ignoring the others, when the classification of this element is finished this element is dismissed and other elements in the input are analyzed.
- The third solution is the temporal correlation approach, first proposed by Malsburg [8], [5]. In this theory, objects belonging to the same entity are bound together in time. In other words, synchronization between different neurons and desynchronization among different regions perform the binding. The classification in our proposed network is based on this third technique. The advantage of this approach is its autonomy, but it is much slower than the first approach.

C. Bio-inspired Neural Networks

Bio-inspired neurons mimic the functional behavior of a real biological neuron. Roughly speaking, these neurones discharge a spike whenever their internal potential exceeds some predefined threshold. In addition, a spike resets the internal potential of the neuron. These neurons are more generalized than classical neural networks. In fact, the information in these networks can be coded in the phase, discharge rate, and the relation between the discharge patterns of the neurons in the network. These codings make bio-inspired neurons mathematically more powerful than classical neural networks [9]. The dynamics of a real biological neuron follows the Hodgkin-Huxley equations. These equations are computationally very expensive. Therefore, some approximations are used to simplify them. The Wang-Terman equations is one of these simplified models. The dynamics of this kind of neurons is governed by a modified version of the Van der Pol relaxation oscillator and follows the following state-space equations, where x is the membrane potential (output) of the neuron and y is the state for channel activation or inactivation.

$$\frac{dx}{dt} = 3x - x^3 + 2 - y + \rho + p + S \quad (1)$$

$$\frac{dy}{dt} = \epsilon[\gamma(1 + \tanh(x/\beta)) - y] \quad (2)$$

ρ denotes the amplitude of a Gaussian noise, p is the external input to the neuron, and S the coupling from other neurons (connections through synaptic weights). ϵ , γ , and β are constants. Since the integration of these equations using the conventional techniques (Runge-Kutta, etc.) is CPU consuming, some simplified techniques have been proposed in the literature to solve the Wang-Terman equations, for instance the algorithmic solution (event-driven) [10] or a technique based on singular limit solutions [11].

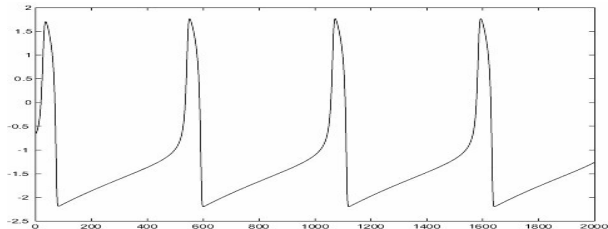


Fig. 1. The output activity of a neuron with $\beta = 100$, $\alpha = 15$, and $\epsilon = 0.1$. The integration method is the fourth order Runge-Kutta and the integration step is set to 0.05

II. SOUND SEGREGATION

A. The preprocessing

The model we use filters vowels (mixed sound sources) by a constant-Q cochlear filter bank. AM demodulation is then applied to higher frequency channels. We are using a filter bank of 24 filters centered on Bark scaled frequencies ranging from 200 Hz to 4.7 kHz. The envelope demodulation (extraction) is done only for channels 5-24. Increasing the number of channels in the filterbank may increase the performance of the separation by reducing "confusion zones" as described in section II.B.

We use the spectrum of the outputs of the cochlear filterbank to create an enhanced version of the Cochleotopic/AMtopic (CAM) map proposed in [12] (for more details about our proposed CAM see [13]). In addition, a reassignment technique is used to enhance the spectra [14]. The CAM is a 2-D representation as shown in Fig. 3. The CAM output for voiced double-vowels has a pseudo-structured pattern. In this paper, we use this pattern to solve the voiced speech segregation problem using our proposed neural architecture. The segregation can be the separation of a voiced speech from another voiced speech or the separation of a voiced speech from an intruding noise.

Supposing that the two sources have different pitches, we can assume that the geometric distance between rays on the map corresponds roughly to the pitch of the underlying source and that this distance is different for different sources. These rays are produced because of the beats between harmonics filtered by the filter belonging to a channel, but are masked by the amplified values of the representation at resonance

frequencies (Fig. 3). This approach lets us enhance rays placed at f_0 , $2f_0$, $3f_0$, etc. on the map, f_0 being the pitch of one of the sources.

B. The Network

The auditory scene analysis in the brain is done in two different stages: segregation and streaming [1]. Segregation consists of finding elementary audio objects in the scene, while streaming is the "binding" of the elements that belong to an object (source). These two steps are implemented in our two-layered neural network.

The first layer is a partially connected network of relaxation oscillators [3]. Each neuron is connected to its four neighbors. The CAM is applied to the input of the neurons. Since the map is sparse, the original 512 points computed for the FFT are down-sampled to 50 points. Therefore, the first layer consists of 24×50 neurons or 1200 neurons. Our observations showed that the geometric interpretation of pitch (ray distance criterion) is less clear for the first four channels. That is why we have also established long-range connections from "clear" (high frequency) zones to "confusion" (low frequency) zones. These connections are only along the "channel number" axis of the CAM (as shown in Fig. 2). This can help the network better extract harmonicity patterns.

The synaptic weights between the neurons of the first layer are adjusted via appropriate adaptation formula based on the difference between inputs (for more details on the adaptation formulas and the dynamic behavior of our proposed architecture see [13]). The second layer is an array of 24 neurons (one for each channel). Each neuron receives the weighted sum of the outputs of the first layer neurons along the frequency axis of the CAM. Since the geometric (Euclidian) distance between rays (spectral maxima) is a function of the pitch of the dominant source in a given channel, the weighted sum of the outputs of the first layers along the frequency axis tells us about the origin of the signal present in that channel. The weights between layer one and layer two are defined as $w_{ll}(f) = \frac{\alpha}{f}$ where f is the frequency bin and α is a constant.

The "binding" of these features is done via this second layer. In fact, the second layer is an array of fully connected neurons along with a global controller. The global controller desynchronizes the synchronized neurons for the first and second sources by emitting inhibitory activities whenever there is an activity (spikings) in the network [3].

III. RESULTS

A mixture of the French /di/ (female speaker) and /da/ (male speaker) (double-vowels) are used to test the system. The signals have equal power, therefore the $SNR = 0dB$. The CAM is extracted for the aforementioned signal. Note that in contrast with most

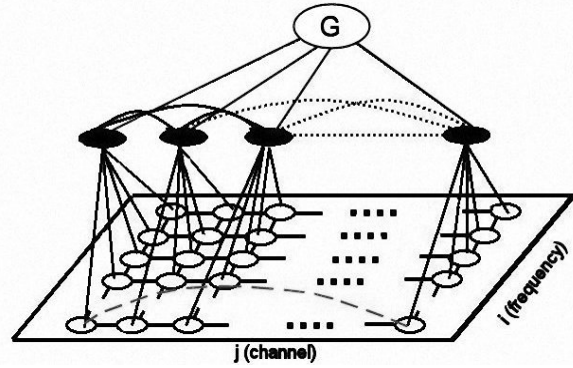


Fig. 2. Architecture of the Two-Layer Bio-inspired Neural Network. G: Stands for global controller (the global controller for the first layer is not shown in the figure). One long range connection is shown in the figure.

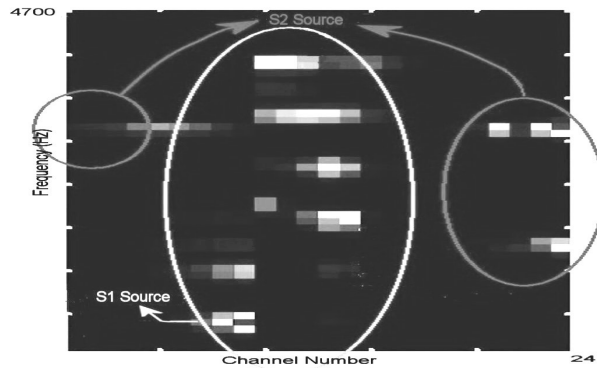


Fig. 3. CAM for the /di/ and /da/ mixture at $SNR = 0$ dB and $t = 166$ ms.

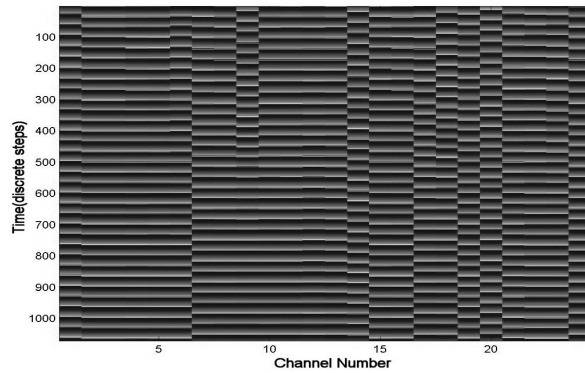


Fig. 4. Spike activity until synchronization for the stimulus presented in Fig. 2 (synchronization time in the order of the number of neurons (24) oscillations).

of the techniques proposed in the literature *no prior pitch detection is made for the sources*. This is in agreement with the physiological observations, which state that no region in the brain is identified as "pitch extractor" and that "pitch extraction" is the byproduct of the Auditory Scene Analysis undertaken in the brain.

Figure 3 shows the CAM for the /di/ and /da/ mixture.

Figure 4 shows the output of the second layer. Note that the binding of channels 1-6 and 19-23 has been made possible through long distance synaptic weights in the second layer. Note that the $H(\cdot)$ (Heaviside function) of the input values are applied to the neurons because of synchronization considerations. Regions with different first layer activity will dissociate through very weak synaptic connections, producing desynchronization (similar frequencies but different phases) and similar region will synchronize (similar frequency and phase) through strong synaptic connections (for more details on performance metrics see [13]).

IV. CONCLUSION AND FURTHER WORK

We proposed a technique to solve the vowel segregation problem using a bio-inspired pre-processing stage and a bio-inspired neural network. We think that the qualitative and quantitative results we obtained from resynthesization are encouraging [13]. In addition we used no prior pitch detector in the approach.

This network only performs a bottom-up processing of speech information. On the other hand, we know that top-down processing is another important aspect of speech segregation. Therefore, two additional layers should compare the output of the first two-layers to saved patterns (i.e., clear vowels). The result of this comparison should then be sent to lower layers, in order to change the dynamics of these inferior levels adequately and to improve the resynthesized sound quality. This approach can be implemented using the Dynamic Link Matching algorithm [15].

Acknowledgments

The authors would like to thank DeLiang Wang for accepting the first author as a visiting student in his laboratory at OSU in August 2001. We would like to thank Romain Balleraud for generating CAM representations for our experiments and for constructive discussions. Many thanks also to NSERC, the UQAC foundation and the University of Sherbrooke for their financial support.

REFERENCES

[1] Al Bregman. *Auditory Scene Analysis*. MIT Press, 1994.
 [2] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, pages 141–177, 2001.
 [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] C. Von der Malsburg. The what and why of binding: The modeler's perspective. *Neuron*, pages 95–104, 1999.
 [6] M. Reisenhuber and T. Poggio. Are cortical models really bound by the binding problem? *Neuron*, 24:87–93, 1999.
 [7] J. Reynolds and R. Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24:19–29, 99.

[8] C. Von der Malsburg and W. Schneider. A neural cocktail-party processor. *Biol. Cybernetics*, pages 29–40, 1986.
 [9] W. Maass and E.D. Sontag. Neural systems as nonlinear Filters. *Neural Computation*, 12(8):1743–1772, 2000.
 [10] D.L. Wang and D. Terman. Image segmentation based on oscillatory correlation. *Neural Computation*, pages 805–836, 1997.
 [11] P.S. Linsay and D. L. Wang. Fast numerical integration of relaxation oscillator networks based on singular limit solutions. *IEEE Trans. on Neural Networks*, pages 523–532, 1998.
 [12] N. Todd. An auditory cortical theory of auditory stream segregation. *Network : Computation in Neural Systems*, 7:349–356, 1996.
 [13] R. Pichevar and J. Rouat. Double-vowel segregation through temporal correlation: A bio-inspired neural network paradigm. In *NOLISP'2003*, 2003.
 [14] F. Plante, G. Meyer, and W. Ainsworth. Improvement of speech spectrogram accuracy by the method of reassignment. *IEEE Trans. on Speech and Audio Processing*, pages 282–287, 1998.
 [15] W. Konen, T. Maurer, and C. Von der Malsburg. A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks*, 7(6/7):1019–1030, 1994.