

# Double-vowel Segregation Through Temporal Correlation: A Bio-inspired Neural Network Paradigm

Ramin Pichevar<sup>1,2</sup> and Jean Rouat<sup>1,2</sup>

(1) Department of Electrical and Computer Engineering, Universit de Sherbrooke, QC, Canada

(2) ERMETIS, Universit du Qubec, Chicoutimi, QC, Canada

emails: pichevar@hermes.usherb.ca, jean.rouat@usherbrooke.ca

**Abstract**—A two-layer spiking neural network is used to segregate double vowels. The first layer is a partially connected spiking neurons of relaxation oscillatory type, while the second layer consists of fully connected relaxation oscillators. A two-dimensional auditory image generated by the enhanced spectrum of cochlear filter bank envelopes is computed. The segregation is based on a channel selection strategy. At each instant of time each channel is assigned to one of the sources present in the auditory scene, i.e. speakers. No prior estimation of pitch for the underlying sources is necessary.

**Index Terms** – Computational Auditory Scene Analysis (CASA), bio-inspired neural networks, sound segregation, cochleotopic/AMtopic map (CAM), temporal binding.

## I. INTRODUCTION

This work deals with the segregation of double vowels using a two-layered spiking neural network. The approach we propose is based on a bio-inspired solution to CASA (Computational Auditory Scene Analysis). Some previous works based on CASA techniques use correlation as a pre-processing stage to segregate sounds or vowels [1], [2], [3]. Another parallel technique to the latter mentioned approach is the use of spectral information [4], [5], [6], [7]. In this work we use a spectral analysis of the filterbank envelopes outputs. Todd [5] proposed a Cochleotopic/AMtopic map, in which a two dimensional representation is generated using two sets of orthogonal filterbanks. In our proposed map, the second dimension of this cochleotopic analysis is replaced by an enhanced FFT analysis (reassigned spectrum [8]). Some references propose expert system solutions to CASA [4], [9] and others neural network based ones [1], [2]. We propose here a bio-inspired approach to the problem that doesn't require any prior detection or knowledge of the pitch of the underlying source signals. In addition, it doesn't compute any correlation.

### A. Computational Auditory Scene Analysis

Humans are able to segregate a desired source in a mixture of sounds ("cocktail-party effect"). This is not the case for computer speech recognition systems. In fact, commercial state of the art speech processing programs work well mostly in quiet environments. The Computational Auditory Scene Analysis (CASA), partially based on psychological and physiological observations, is aimed to solve that kind of shortages.

### B. Bio-inspired Neural Networks

Bio-inspired neural networks are dynamic. In fact, information in this kind of networks is coded in the spike timing of neurons, which is a generalization of the classical neural networks. It is possible to design more autonomous and flexible neural networks using these bio-inspired neurons than by using conventional ones. In this work, we use a two-layered bio-inspired neural network.

The building blocks of this network are oscillatory neurons [1]. The dynamics of this kind of neurons is governed by a modified version of the Van der Pol relaxation oscillator (called the Wang-Terman oscillator) as described in section II-B. There is an active phase when the neuron spikes and a relaxation phase when the neuron is silent.

### C. The Binding Problem

We can assume that our segregation problem is a generalized classification problem, in which we want to bind features to different sources. This generalized classification problem was first addressed by Rosenblatt [10].

Temporal correlation theory was first proposed by Malsburg [11], [10] to solve the binding problem. 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 [12].

## II. THE ARCHITECTURE

### A. The preprocessing

The CAM output for voiced double-vowels has a pseudo-structured pattern.

We use 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 [13].

We use the spectrum of the outputs of the cochlear filterbank to create an enhanced version of the Cochleotopic/AMtopic (CAM) map proposed in [5]. Our CAM generation algorithm is as follows. The sampling rate of the signal is 16000 samples/s.

- Extract the envelope (AM demodulation) for channels 5-24; for other channels use raw outputs.

- Compute the STFT (Short Time Fourier Transform) using a 1024 Hamming window (equal to 64 ms).
- In order to increase the spectro-temporal resolution of the STFT, find the reassigned spectrum of the STFT [8] (this consists of applying an affine transform to the points in order to relocate the spectrum).
- Compute the logarithm of the result. This stage enhances the pattern of the underlying harmonicity we want to extract.

Another approach proposed in the literature is to replace the CAM by correlograms [1], [3]. Autocorrelation-based models rely on the observation that, while a representation based on the average discharge rate in an auditory filterbank is unable to resolve speech harmonics in the high-frequency range directly, precise temporal information, extracted by correlograms, is present in the discharge pattern seen in each channel.

For voiced sound, glottal impulses are convolved with the transfer function of the vocal tract. 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 by 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. 2). 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 [14]. 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 dynamics of the neurons follows the following state-space equations, where  $x_i$  is the membrane potential (output) of the neuron and  $y_i$  is the state for channel activation or inactivation.

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

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

$\rho$  denotes the amplitude of a Gaussian noise,  $p$  the input to the neuron, and  $S$  the coupling from other neurons (connections through synaptic weights).  $\epsilon$ ,  $\gamma$ , and  $\beta$  are constants. Initial values are generated by a uniform distribution between the interval  $[-2; 2]$  for  $x$  and between  $[0; 8]$  for  $y$  (these values correspond to the whole dynamic range of the equations). Forward Euler integration with a step size of 0.01 is used to solve equations 1 and 2. Bigger integration step sizes will lead to complex network behaviors such as antiphase synchrony or loose synchrony [15].

The first layer is a partially connected network of relaxation oscillators [1]. 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  $24 \times 50$  neurons or 1200 neurons. Our observations showed that the geometric interpretation of pitch (ray distance criterion) is less clear for the first four channels. For this reason, we have also established long-range connections from "clear" (high frequency) zones to "confusion" (low frequency) zones. These connections are defined only across the "channel number" axis of the CAM. This can help the network better extract harmonicity patterns.

The layer can be reset by a master neuron that acts as a master clock [16]. This clock can reset the network, so that it doesn't remember the long past.

The synaptic weight between  $neuron(i, j)$  and  $neuron(k, m)$  of the first layer is computed via the following formula:

$$w_{i,j,k,m}(t) = \frac{1}{Card\{N(i, j)\}} \frac{0.25}{e^{\lambda|p(i,j;t)-p(k,m;t)|}} \quad (3)$$

here  $p(i, j)$  and  $p(k, m)$  are respectively external inputs to  $neuron(i, j)$  and  $neuron(k, m) \in N(i, j)$ .  $Card\{N(i, j)\}$  is a normalization factor and is equal to the cardinal number (number of elements) of the set  $N(i, j)$  containing neighbors connected to the  $neuron(i, j)$  (can be equal to 4, 3 or 2 depending on the location of the neuron on the map, i.e. center, corner, etc. and whether the weight between the neuron and its neighbors is greater than a predefined threshold). The external input values are normalized. The value of  $\lambda$  depends on the dynamic range of the inputs and is set to  $\lambda = 1$  in our case. This same weight adaptation is used for "long range clear to confusion zone" connections (Eq. 5). The influence  $S_{i,j}$  defined in Eq. 1 is computed by :

$$S_{i,j}(t) = \sum_{k,m \in N(i,j)} w_{i,j,k,m}(t)H(x(k, m; t)) - G(t) + L_{i,j}(t) \quad (4)$$

$H(\cdot)$  is the Heaviside function,  $G(t)$  is the influence of the global controller as defined in [1], and  $L_{i,j}(t)$  is the long range influence:

$$L_{i,j}(t) = \begin{cases} 0 & j > 4 \\ \sum_{k=14,15,23,24} w_{i,j,i,k}(t)H(x(i, k; t)) & j < 4 \end{cases} \quad (5)$$

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  $w_{ll}(i) = \frac{\alpha}{i}$  where  $i$  can be related to the frequency bins and  $\alpha$  is a constant. Therefore the input stimulus to the neuron( $j$ ) in the second layer is defined as follows:

$$\theta(j; t) = \sum_i w_{ll}(i) \overline{x(i, j; t)} \quad (6)$$

Where  $\overline{x(i, j; t)}$  is the output of the first layer for channel  $j$ , at time  $t$ , and for frequency  $i$ , averaged over a time window (the

length of the window is in the order of the discharge period).  $\theta(j; t)$  is the input to the neuron  $j$  in the second layer at time  $t$ . The synaptic weights in the second layer are adjusted through the following rule:

$$w'_{ij}(t) = \frac{0.2}{e^{\mu|p(j;t)-p(k;t)|}} \quad (7)$$

$\mu$  is chosen to be equal to 2. 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 [1].

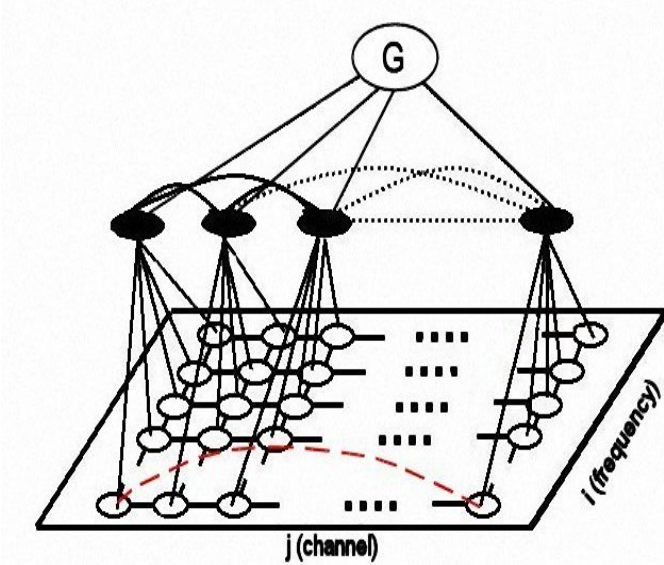


Fig. 1. Architecture of the Two-Layer Bio-inspired Neural Network. G: Stands for global controller (the global controller for the first layer is not shown on the figure). One long range connection is shown in the figure.

Although for the given double vowel separation problem, the CAM doesn't vary so much in time, in the case of unvoiced speech (like stop consonants, etc.) or fast changing noises the temporal aspect could become very important. Hence, this architecture should be also useful for that kind of problems.

### 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 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 2 shows the CAM for the /di/ and /da/ mixture.

Figure 3 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. Since there is no energy (or very little energy) in channels

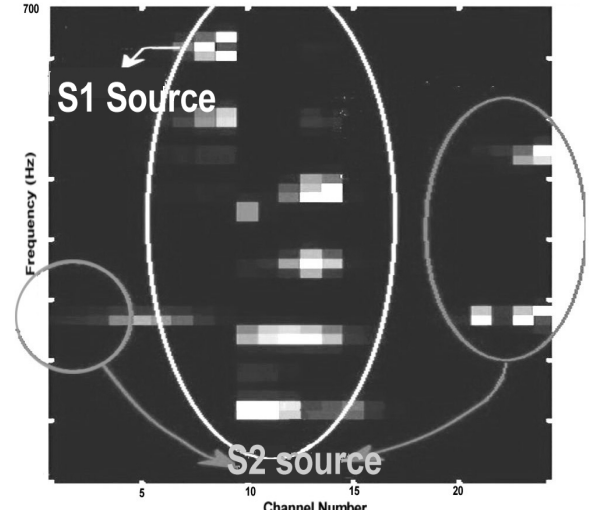


Fig. 2. CAM for the /di/ and /da/ mixture at  $SNR = 0$  dB and  $t = 166$  ms.

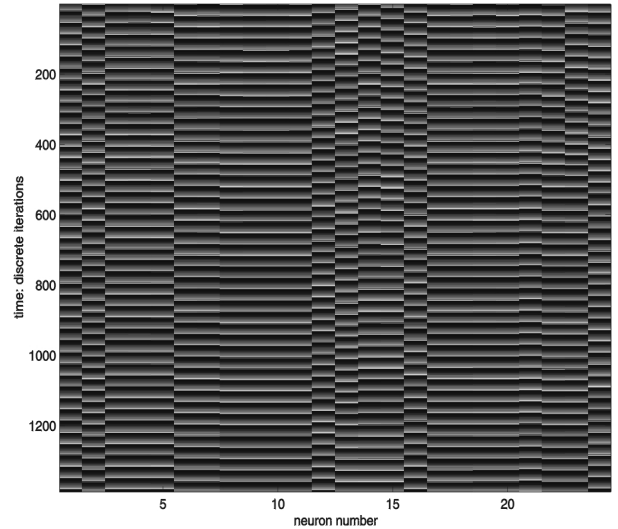


Fig. 3. Spike activity until synchronization for the stimulus presented in Fig. 2 (synchronization time in the order of the number of neurons (24) oscillations).

12-16, the network has bound those channels arbitrarily to the first or to the second source. There could be a slight difference of frequency between different synchronized zones if intensity levels are directly applied as input. This could lead to what is called "partial synchronization". In order to circumvent this problem, we decided to apply the  $H(\cdot)$  of the input to the neurons so that the intensity of all stimuli is equal. The initial intensity difference between regions is implicitly applied through synaptic connections. 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.

We use the PEL (Percentage of Energy Loss) criterion to measure the performance of our system. The PEL is defined

as follows:

$$PEL = \frac{\sum_t e^2(t)}{\sum_t O^2(t)} \quad (8)$$

Where  $e(t)$  is the difference between the desired output and the actual resynthesized output  $O(t)$ .

The PEL for the synthesized /da/ is 24.69% at SNR= 0dB and is equal to 29.72% for the /di/. Perceptual tests have shown that although we lose some sound quality after the process, the vowels are separated and sound is recognizable. This is an important aspect in "speech enhancement", because some methods may have good PELs but fair perceptive quality (discontinuities in the resynthesized speech, etc.).

#### IV. CONCLUSION AND FURTHER WORK

We proposed a technique to solve the double-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 resynthesis demonstrate a strong potential for the approach. In addition we used no prior pitch detector in the architecture.

The lack of resolution in low-frequency channels makes the binding process difficult in those channels and becomes a source of energy loss in the resynthesized output. More work should be done to increase the resolution. This could be done by an increase in the number of channels, a modification in the CAM computation, or even by changing the network architecture. Increasing the number of channels in the filterbank may increase the performance of the separation by reducing "confusion zones" (described in section II-B). Also, replacing cochlear filters by non-overlapping ones should help us decrease the confusion zone and gives us a better resynthesis performance. Preliminary results also show that a similar map can be generated using the Instantaneous Frequency (IF) using FM demodulation techniques.

By reducing the hamming window length, one should be able to detect onset/offset times for consonants, which will be popped out in the CAM representation. The network would be able to bind the information using offset/onset of information bins.

Top-down (schema-driven) processing of information can also be used to further enhance the segregation by matching pre-stored patterns to the signal extracted by the bottom-up method proposed in this article. Oscillatory Dynamic Link Matching can be used for this purpose [17].

In addition to the CAM, other representations can be used for other types of signals. This multiple representation strategy can be seen as a top-down (schema-driven) processing. In fact efferent connections from higher auditory levels could control the stiffness of cochlear hair cells, giving birth to different auditory maps [18].

#### Acknowledgments

The authors would like to thank DeLiang Wang for fruitful discussions on oscillatory neural networks. Also, we would like to thank Romain Balleraud for generating CAM representations for our experiments and for constructive discussions. NSERC, the UQAC foundation, and the University of Sherbrooke supported us financially.

#### REFERENCES

- [1] 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.
- [2] J. Rouat and R. Pichevar. Nonlinear speech processing techniques for source segregation. In *EUSIPCO, Toulouse, France, 2002*.
- [3] M. J. Hewitt R. Meddis. Modelling the identification of concurrent vowels with different fundamental frequencies. *JASA*, 91:224–233, 1992.
- [4] G. Meyer, D. Yang, and W. Ainsworth. Applying a model of concurrent vowel segregation to real speech. *Computational models of auditory function*, pages 297–310, 2001.
- [5] N. Todd. An auditory cortical theory of auditory stream segregation. *Network : Computation in Neural Systems*, 7:349–356, 1996.
- [6] J. Tchorz and B. Kollmeier. SNR estimation based on amplitude modulation analysis with applications to noise suppression. *IEEE Trans. on Speech and Audio Processing*, 11:184–192, may 2003.
- [7] M. Casey. Separation of mixed audio sources by independent subspace analysis. In *Int'l Computer Music Conference, Berlin, Germany, 2000*.
- [8] 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.
- [9] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Comm.*, pages 141–177, 2001.
- [10] C. Von der Malsburg. The what and why of binding: The modeler's perspective. *Neuron*, pages 95–104, 1999.
- [11] C. Von der Marlsburg and W. Schneider. A neural cocktail-party processor. *Biol. Cybernetics*, pages 29–40, 1986.
- [12] R. Pichevar and J. Rouat. Binding of audio elements in the sound source segregation problem via a two-layered bio-inspired neural network. In *IEEE CCECE, Montreal, Canada, 2003*.
- [13] J. Rouat, Y. C. Liu, and D. Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Comm.*, 21:191–207, 1997.
- [14] Al Bregman. *Auditory Scene Analysis*. MIT Press, 1990.
- [15] S. R. Campbell and D. Wang. Relaxation oscillators with time delay coupling. *Physica D*, pages 151–178, 1998.
- [16] R. Sarpeshkar and M. O'Halloran. Scalable hybrid computation with spike. *Neural Computation*, pages 2003–2038, 2002.
- [17] R. Pichevar and J. Rouat. Oscillatory dynamic link matching for pattern recognition. In *International Workshop on Neural Coding (NCWS), Aulla, Italy, 2003*.
- [18] R. Pichevar and Jean Rouat. Cochleotopic/AMtopic (CAM) and Cochleotopic/Spectrotopic (CSM) map based sound source separation using relaxation oscillatory neurons. In *IEEE Neural Networks for Signal Processing Workshop, Toulouse, France, 2003*.