

COCHLEOTOPIC/AMTOPIC (CAM) AND COCHLEOTOPIC/SPECTROTOPIC (CSM) MAP BASED SOUND SOURCE SEPARATION USING RELAXATION OSCILLATORY NEURONS

Ramin Pichevar and Jean Rouat

Department of Computer Eng., Université de Sherbrooke, QC, Canada

ERMETIS, Université du Québec, Chicoutimi, QC, Canada

emails: pichevar@hermes.usherb.ca, jean.rouat@ieee.org

Abstract. We use a two-layered unsupervised bio-inspired neural network to segregate sound sources, e.g. double-vowels or vowels intruded by nonstationary noise sources. 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. We show that in order to correctly segregate sound sources, we should either use Cochleotopic/AMtopic Map (CAM) or Cochleotopic/Spectrotopic Map (CSM) depending on the nature of the intruding sound source.

INTRODUCTION

Motivation

We explore the potential of unsupervised monophonic source separation with no prior knowledge of the underlying sound signals. We study a bio-inspired solution, in which pseudo auditory images are obtained from two different representations (Cochleotopic/AMtopic Map and Cochleotopic/Spectrotopic Map). These auditory images are then processed further by a two-layered neural network. The first layer segments the sound into auditory objects while the second layer binds objects that belong to the same source.

How well can we separate sources without any complex post-processing of the representations (CAM, CSM), with no prior knowledge (explicit or statistical) of the underlying signals, and by using our proposed neural network architecture? Is it possible to design such a system? What would be the limitations for practical applications?

Most monophonic source separation systems are based on either expert systems [3] (explicit knowledge), or on statistical approaches [7] [14] (implicit knowledge), or on bio-inspired approaches [4] [13] [15]. Jang and Lee [7], and Roweis [14] have proposed extensions of data driven methods to the problem of monophonic source separation. Wang and Brown [15] have proposed an original approach that uses features obtained from correlograms, estimates F0 (the pitch), and uses an oscillatory neural network. Our system neither needs a prior knowledge of the underlying sources, nor it estimates F0, nor it computes the computationally expensive correlograms. The architecture is also designed to handle continuous input signals (no need to segment the signal into time frames) and is based on the availability of simultaneous auditory representations of signals.

It is known that the peripheral auditory system adaptively extracts representations suitable for higher auditory nucleus processing. Furthermore, tonotopic maps are observed in the colliculus and specialized cells in the cochlear nucleus. We therefore infer that multiple representations of the same signal are available to the auditory centers and we propose to build a source separation system that can simultaneously use two of these representations.

It has also been recently observed that the efferent loop between the medial olivocochlear system (MOC) and the outer hair cells modifies the cochlear response in such a way that speech is enhanced from the background noise [8]. Furthermore, the peripheral auditory system can enhance the AM, the FM, the envelope, the transient components of the signal, etc. We suppose here that envelope detection and selection between the CAM and the CSM, in the auditory pathway, could be associated with the change of stiffness of hair cells combined with cochlear nucleus cell processing [6] [9].

Binding of Auditory Sources

We assume that sound segregation is a generalized classification problem, in which we want to bind features extracted in different sections of our neural network map.

Three solutions to the binding (or to circumvent the binding) are proposed in the literature: hierarchical coding, temporal correlation, and attentional models. We use the temporal correlation approach [5], in which 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 advantage of this approach is its autonomy, but it is much slower than the hierarchical approach.

Bio-inspired Neural Networks

Bio-inspired neurons mimic the functional behavior of real biological neurons. In fact, the information in these bio-inspired networks can be coded in the spike phase, in the spike discharge rate, and into the relation between the discharge patterns of the neurons in the network.

SOUND SEGREGATION

The preprocessing

Two types of front-end processing are used for sound source segregation. The selection between these two representations for a given auditory scene could be seen as a top-down feedback. In fact, as stated earlier, the difference between these two maps is that envelope detection is done for the CAM but not for the CSM. In the *results* section, we show that each representation is suitable for a specific class of intruding sources. For both maps, the mixed sound source is filtered by a constant-Q cochlear filter bank with linear phase [12].

In the double-vowel segregation case, we are looking for structured patterns in the channels. In fact, for voiced speech we base our segregation algorithm on the hypothesis that two speakers do not have the same pitch. Thus, the outputs of the filterbank are AM demodulated (envelope detection) at least for higher channels and then the reassigned spectrum STFT (Short Time Fourier Transform) is computed [11]. The magnitude of the STFT is a structured pattern where the distance between the rays on the map is a function of the pitch of the signal. 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 generated by the beats between harmonics present in a channel of the cochlear filterbank (fig. 3). On the map, this approach lets us enhance rays placed at f_0 , $2f_0$, $3f_0$, etc. f_0 is the pitch of one of the sources.

In the voice plus siren case, a highly nonstationary noise source is mixed with the vowel. At each instant, frequency tone bursts are filtered by the filterbank. Note that due to nonstationarity these bursts are short in time. Thus, we are looking for short but high energy bursts. Envelope detection in this case will fade this phenomenon. Hence, Cochleotopic/ Spectrotopic Map (CSM) is generated.

Our CAM/CSM generation algorithm is as follows. The signal is down-sampled to 12000 samples/s.

- Filter the sound source using a 24-filter bark-scaled cochlear filterbank ranging from 200 Hz to 4.7 kHz .
- For CAM: Extract the envelope (AM demodulation) for channels 5-24; for other channels use raw outputs [12].
For CSM: Nothing is done in this step.
- Compute the STFT using a Hamming window.
- In order to increase the spectro-temporal resolution of the STFT, find

the reassigned spectrum of the STFT [11] (this consists of applying an affine transform to the points in order to relocate the spectrum).

- Compute the logarithm of the magnitude of the STFT. The logarithm enhances the presence of the stronger source in a given 2D frequency bin of the CAM/CSM ¹.

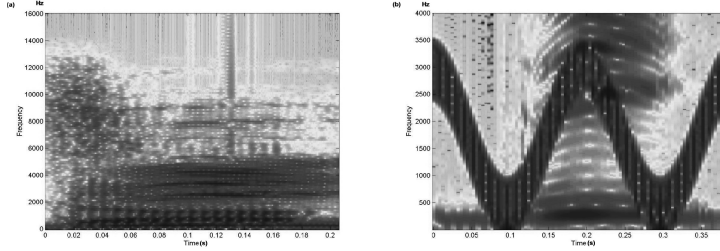


Figure 1: (a) Spectrogram of the /di/ and /da/ mixture. (b) Spectrogram of the /di/ plus siren mixture

The Neural Network

The dynamics of the neurons we use is governed by a modified version of the Van der Pol relaxation oscillator (Wang-Terman oscillators [15]). The state-space equations for this dynamics are as follows:

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

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

Where x is the membrane potential (output) of the neuron and y is the state for channel activation or inactivation. ρ denotes the amplitude of a Gaussian noise, p is the external input to the neuron, and S is the coupling from other neurons (connections through synaptic weights). ϵ , γ , and β are constants. The first layer is a partially connected network of relaxation oscillators [15]. Each neuron is connected to its four neighbors. The CAM (or the CSM) 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 (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 exist only across the "cochlear channel number" axis of the CAM. This architecture can help the network better extract harmonic patterns.

¹ $\log(e_1 + e_2) \simeq \max(\log e_1, \log e_2)$ (unless e_1 and e_2 are both large and almost equal) [14]

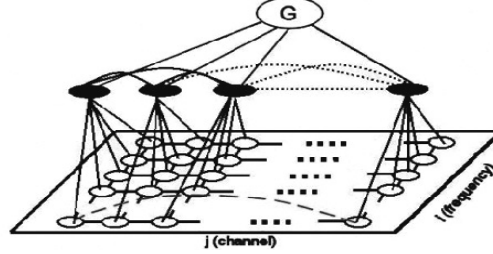


Figure 2: Architecture of the Two-Layer Bio-inspired Neural Network. G: Stands for global controller (the global controller for the first layer is not shown on the figure). One long range connection is shown in the figure.

The layer can be reset by a master neuron that acts as a master clock so that the network doesn't remember the long-term 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.). The external input values are normalized. The value of λ 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. 8) in CAM processing case. The coupling $S_{i,j}$ defined in Eq. 1 is defined as :

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

$H(\cdot)$ is the Heaviside function, the dynamics of $G(t)$ (the global controller) is as follows:

$$G(t) = \alpha H(z - \theta) \quad (5)$$

$$\frac{dz}{dt} = \sigma - \xi z \quad (6)$$

σ is equal to 1 if the global activity of the network is greater than a predefined ζ and is zero otherwise. α and ξ are constants.

$L_{i,j}(t)$ is the long range coupling as follows:

$$L_{i,j}(t) = \begin{cases} 0 & j > 4 \\ \sum_{k=14,15,23,24} w_{i,j,i,k}(t) H(x(i, k; t)) & j < 4 \end{cases} \quad (7)$$

κ is a binary variable defined as follows:

$$\kappa = \begin{cases} 1 & \text{for } CAM \\ 0 & \text{for } CSM \end{cases} \quad (8)$$

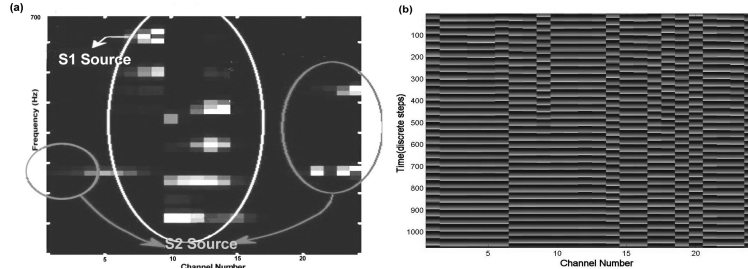


Figure 3: (a) CAM for the female /di/ and male /da/ mixture at $SNR = 0$ dB and $t = 166$ ms. (b) Spike activity until synchronization for the stimulus presented in (a) (synchronization time in the order of the number of neurons (24) oscillations)

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/CSM.

- For 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.
- For the CSM: Highly localized energy bursts will be enhanced by that representation.

The weights between layer one and layer two are defined as $w_{li}(i) = \frac{\alpha}{i}$, where i can be related to the frequency bins of the STFT and α is a constant. Therefore the input stimulus to $neuron(j)$ in the second layer is defined as follows:

$$\theta(j; t) = \sum_i w_{li}(i) \overline{x(i, j; t)} \quad (9)$$

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 duration 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 inside the second layer are adjusted through the following rule:

$$w'_{ij}(t) = \frac{0.2}{e^{\mu|p(j;t)-p(k;t)|}} \quad (10)$$

μ 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 [15].

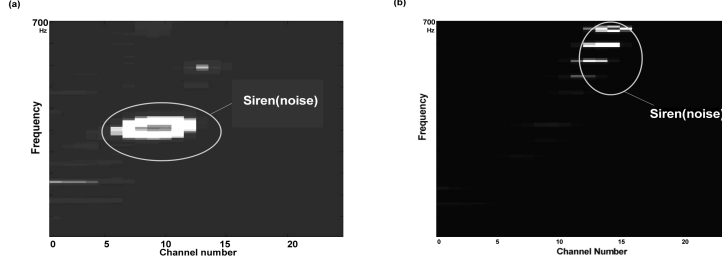


Figure 4: CSM of the mixture of /di/ and the siren in Eq. 12 at (a) $t=50$ ms (b) $t=200$ ms

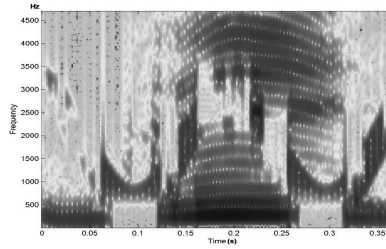


Figure 5: The spectrogram of the siren plus speech sound after processing by our proposed neural network using the CSM

PRELIMINARY RESULTS

Segregation with the CAM

A mixture of the French /di/ (female speaker) and /da/ (male speaker) (double-vowels) is used to test the system. The signals have equal power, therefore the $SNR = 0dB$ (Fig. 1 (b)). 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*.

Fig. 3 (a) shows the CAM for the /di/ and /da/ mixture. Fig. 3(b) 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 also, 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.

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

$$PEL = \frac{\sum_t e^2(t)}{\sum_t O^2(t)} \quad (11)$$

Where $e(t)$ is the difference between the desired output and the actual syn-

thesized 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 PEL but fair perceptive quality.

Segregation with the CSM

A modified version of the siren used in Cooke's database [2] [15] (Eq. 12) is mixed with the /di/ vowel. The spectrogram of the mixed sound is shown in Fig. 1 (b). The noise is represented by the following equation and can be generated by a VCO (Voltage controlled oscillator):

$$n(t) = \sum_i \cos\left[\left(\omega_i t + \frac{\Delta\omega}{\omega_m} \cos(\omega_m t + \varphi_i)\right)\right] \quad (12)$$

Where ω_i is the central angular frequency, ω_m is the angular frequency of the modulating signal, $\Delta\omega$ is the angular frequency deviation, and φ_i is the phase of the modulating signal (equal to 0 in Fig. 1 (b)). As in the previous case (CAM), the sum of the output of the first layer along the frequency axis is different when the siren is present. The binding of channels on the two sides of the "noise intruding zone" is done via the long-range synaptic connections of the second layer. The spectrogram of the result is shown in Fig. 5. A CSM is extracted at each 10 ms and the selection is made by 10 ms intervals. In a future work, we will use much smaller selection intervals and shorter STFT windows to prevent discontinuities in Fig. 5. Furthermore, overlapping cochlear filters are not suitable for the synthesis of the processed speech.

We use here, a siren noise similar to the one proposed by Cooke [2]. If we were using two sirens with opposite phases ($\varphi_i = \frac{\pi}{2}$ and $\varphi_i = 0$), then two distinct active regions from the network would have popped out at each instant of time and a binding mechanism like the one we propose should be used to segregate sound sources. The details of such an experiment will be given in a further work. The results can be heard and evaluated at [1].

Mask and synthesis

The synthesis is performed as follows:

$$s(t) = \sum_{i=1}^{24} m_i(t) z_i(t) \quad (13)$$

where $s(t)$ is the recovered signal, $z_i(t)$ is the filtered output of the original corrupted signal for channel i and $m_i(t)$ is the mask value. The mask has equal values for all channels whose associated neurons are synchronized, e.g. $m_i(t) = 0$ or 1, depending on the source to be enhanced.

Acknowledgments

The authors would like to thank DeLiang Wang for fruitful discussions on relaxation neural networks. Romain Balleraud generated the CAM representations for our experiments. Many thanks to the three anonymous reviewers for helpful comments, criticisms and references. NSERC, UQAC, and the University of Sherbrooke supported us financially.

REFERENCES

- [1] <http://www-edu.gel.usherbrooke.ca/pichevar/>.
- [2] M. Cooke, <http://www.dcs.shef.ac.uk/~martin/>.
- [3] M. Cooke and D. Ellis, "The Auditory Organization of Speech and Other Sources in Listeners and Computational Models," **Speech Comm.**, pp. 141–177, 2001.
- [4] M. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," **Speech Communication**, vol. 34, pp. 267–285, 2001.
- [5] C. V. der Marlsburg and W. Schneider, "A Neural Cocktail-Party Processor," **Biol. Cybernetics**, pp. 29–40, 1986.
- [6] C. Giguere and P. C. Woodland, "A Computational Model of the Auditory Periphery for Speech and Hearing Research," **JASA**, pp. 331–349, 1994.
- [7] G. Jang and T. Lee, "A Maximum Likelihood Approach to Single-Channel Source Separation," **Signal Processing Letters**, pp. 168–171, 2003.
- [8] S. Kim, D. R. Frisina and R. D. Frisina, "Effects of Age on Contralateral Suppression of Distorsion Product Otoacoustic Emissions in Human Listeners with Normal Hearing," **Audiology Neuro Otology**, pp. 7:348–357, 2002.
- [9] M. Liberman, S. Puria and J. J. Guinan, "The ipsilaterally evoked olivocochlearreflex causes rapid adaptation of the 2f1-f2 distortion product otoacoustic emission," **JASA**, vol. 99, pp. 2572–3584, 1996.
- [10] R. Pichevar and J. Rouat, "Oscillatory Dynamic Link Matching for Pattern Recognition," in **International Workshop on Neural Coding (NCWS), Aulla, Italy**, 2003.
- [11] F. Plante, G. Meyer and W. Ainsworth, "Improvement of Speech Spectrogram Accuracy by the Method of Reassignment," **IEEE Trans. on Speech and Audio Processing**, pp. 282–287, 1998.
- [12] J. Rouat, Y. C. Liu and D. Morissette, "A Pitch Determination and Voiced/Unvoiced Decision Algorithm for Noisy Speech," **Speech Comm.**, vol. 21, pp. 191–207, 1997.
- [13] J. Rouat and R. Pichevar, "Nonlinear Speech Processing Techniques for Source Segregation," in **EUSIPCO, Toulouse, France**, 2002.
- [14] S. T. Roweis, "One Microphone Source Separation," in **NIPS, Denver, USA**, 2000.
- [15] D. Wang and G. J. Brown, "Separation of Speech from Interfering Sounds Based on Oscillatory Correlation," **IEEE Transactions on Neural Networks**, vol. 10, no. 3, pp. 684–697, May 1999.