EURASIP Journal on Applied Signal Processing 2005:9, 1365–1373

Source Separation with One Ear: Proposition for an Anthropomorphic Approach

Jean Rouat

Département de Génie Électrique et de Génie Informatique, Université Sherbrooke, 2500 boulevard de l'Université, Sherbrooke, QC, Canada J1K 2R1

Équipe de Recherche en Micro-électronique et Traitement Informatique des Signaux (ETMETIS), Département de Sciences Appliqués, Université du Québec à Chicoutimi, 555 boulevard de l'Université, Chicoutimi, Québec, Canada G7H 2B1
Email: jean.rouat@ieee.org

Ramin Pichevar

Département de Génie Électrique et de Génie Informatique, Université Sherbrooke, 2500 boulevard de l'Université, Sherbrooke, QC, Canada J1K 2R1
Email: ramin.pichevar@usherbrooke.ca

Équipe de Recherche en Micro-électronique et Traitement Informatique des Signaux (ETMETIS), Département de Sciences Appliqués, Université du Québec à Chicoutimi, 555 boulevard de l'Université, Chicoutimi, Québec, Canada G7H 2B1

Received 9 December 2003; Revised 23 August 2004

We present an example of an anthropomorphic approach, in which auditory-based cues are combined with temporal correlation to implement a source separation system. The auditory features are based on spectral amplitude modulation and energy information obtained through 256 cochlear filters. Segmentation and binding of auditory objects are performed with a two-layered spiking neural network. The first layer performs the segmentation of the auditory images into objects, while the second layer binds the auditory objects belonging to the same source. The binding is further used to generate a mask (binary gain) to suppress the undesired sources from the original signal. Results are presented for a double-voiced (2 speakers) speech segment and for sentences corrupted with different noise sources. Comparative results are also given using PESQ (perceptual evaluation of speech quality) scores. The spiking neural network is fully adaptive and unsupervised.

Keywords and phrases: auditory modeling, source separation, amplitude modulation, auditory scene analysis, spiking neurons, temporal correlation.

1. INTRODUCTION

1.1. Source separation

Source separation of mixed signals is an important problem with many applications in the context of audio processing. It can be used to assist robots in segregating multiple speakers, to ease the automatic transcription of videos via the audio tracks, to segregate musical instruments before automatic transcription, to clean up signal before performing speech recognition, and so forth. The ideal instrumental setup is based on the use of arrays of microphones during recording to obtain many audio channels.

In many situations, only one channel is available to the audio engineer that still has to solve the separation problem. Most monophonic source separation systems require *a priori* knowledge, that is, expert systems (explicit knowledge) or statistical approaches (implicit knowledge) [1]. Most of these systems perform reasonably well only on specific signals (generally voiced speech or harmonic music) and fail to efficiently segregate a broad range of signals. Sameti [2] uses hidden Markov models, while Roweis [3, 4] and Royes-Gomez [5] use factorial hidden Markov models. Jang and Lee [6] use maximum a posteriori (MAP) estimation. They all require training on huge signal databases to estimate probability models. Wang and Brown [7] have first proposed an original bio-inspired approach that uses features obtained from correlograms and F0 (pitch frequency) in combination with an oscillatory neural network. Hu and Wang use a pitch tracking technique [8] to segregate harmonic sources. Both systems are limited to harmonic signals.

We propose here to extend the bio-inspired approach to more general situations without training or prior knowledge of underlying signal properties.

1.2. System overview

Physiology, psychoacoustic, and signal processing are integrated to design a multiple-source separation system when

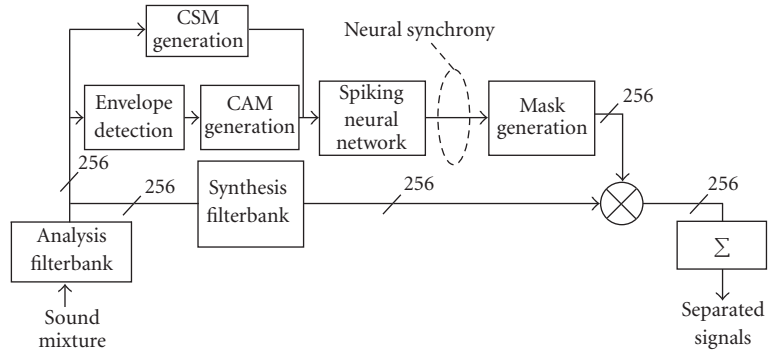

Figure 1: Source separation system. Depending on the sources' auditory images (CAM or CSM), the spiking neural network generates the mask (binary gain) to switch on/off—in time and across channels—the synthesis filter bank channels before final summation.

only one audio channel is available (Figure 1). It combines a spiking neural network with a reconstruction analysis/synthesis cochlear filter bank along with auditory image representations of audible signals. The segregation and binding of the auditory objects (coming from different sound sources) is performed by the spiking neural network (implementing the *temporal correlation* [9, 10]) that also generates a mask[1] to be used in conjunction with the synthesis filter bank to generate the separated sound sources.

The neural network uses third-generation neural networks, where neurons are usually called *spiking* neurons [11]. In our implementation, neurons firing at the same instants (same firing phase) are characteristic of similar stimuli or comparable input signals.[2] Usually *spiking* neurons, in opposition to *formal* neurons, have a constant firing amplitude. This coding yields noise and interference robustness while facilitating adaptive and dynamic synapses (link between neurons) for unsupervised and autonomous system design. Numerous spike timing coding schemes are possible (and observable in physiology) [12]. Among them, we decided to use synchronization and oscillatory coding schemes in combination with a competitive unsupervised framework (obtained with dynamic synapses), where groups of synchronous neurons are observed. This choice has the advantage to allow the design of unsupervised systems with no training (or learning) phase. To some extent, the neural network can be viewed as a map where links between neurons are dynamic. In our implementation of the *temporal correlation*, two neurons with similar inputs on their dendrites will increase their soma to soma synaptic weights (dynamic synapses), forcing synchronous response. On the other hand, neurons with dissimilar dendritic inputs will have reduced soma to soma synaptic weights yielding reduced coupling and asynchronous neural responses.

[1]Mask and masking refer here to a binary gain and should not be confused with the conventional definition of masking in psychoacoustics.

[2]The information is coded in the firing instants.

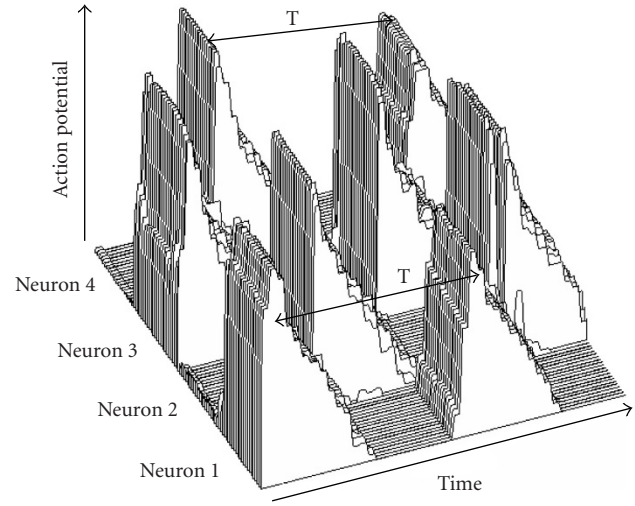

Figure 2: Dynamic temporal correlation for two simultaneous sources: time evolution of the electrical output potential for four neurons from the second layer (output layer). T is the oscillatory period. Two sets of synchronous neurons appear (neurons 1 and 3 for source 1; neurons 2 and 4 for source 2). Plot degradations are due to JPEG coding.

Figure 2 illustrates the oscillatory response behavior of the output layer of the proposed neural network for two sources.

Compared to conventional approaches, our system does not require a priori knowledge, is not limited to harmonic signals, does not require training, and does not need pitch extraction. 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. Our approach is inspired by knowledge in anthropomorphic systems but is not an attempt to reproduce physiology or psychoacoustics.

The next two sections motivate the anthropomorphic approach, Section 4 describes in detail the system, Section 5 describes the experiments, Section 6 gives the results, and Section 7 is the discussion and conclusion.

2. ANTHROPOMORPHIC APPROACH

2.1. Physiology: multiple features

Schreiner and Langner in [13, 14] have shown that the inferior colliculus of the cat contains a highly systematic topographic representation of AM parameters. Maps showing best modulation frequency have been determined. The pioneering work by Robles et al. in [15, 16, 17] reveals the importance of AM-FM[3] coding in the peripheral auditory system along with the role of the efferent system in relation to adaptive tuning of the cochlea. In this paper, we use energy-based features (Cochleotopic/Spectrotopic Map) and AM features (Cochleotopic/AMtopic Map) as signal representations. The proposed architecture is not limited by the number of representations. For now, we use two representations to illustrate the relevance of multiple representations of the signal available along the auditory pathway. In fact, it is clear from physiology that multiple and simultaneous representations of the same input signal are observed in the cochlear nucleus [18, 19]. In the remaining parts of the paper, we call these representations *auditory images*.

2.2. Cocktail-party effect and CASA

Humans are able to segregate a desired source in a mixture of sounds (*cocktail-party effect*). Psychoacoustical experiments have shown that although binaural audition may help to improve 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 (e.g., when someone listens to a radio broadcast) [20]. Using the knowledge acquired in visual scene analysis and by making an analogy between vision and audition, Bregman developed the key notions of the *auditory scene analysis* (ASA) [20]. Two of the most important aspects in ASA are the *segregation* and *grouping (or integration)* of sound sources. The segregation step partitions the auditory scene into fundamental auditory elements and the grouping is the binding of these elements in order to reproduce the initial sound sources. These two stages are influenced by top-down processing (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 characteristics [7, 21]. For a review see [1].

2.3. Binding of auditory sources

We assume here that sound segregation is a generalized classification problem in which we want to bind features extracted from the auditory image representations in different regions of our neural network map. We use the temporal correlation approach as suggested by Milner [9] and Malsburg in [22, 23] who observed that synchrony is a crucial feature to bind neurons associated to similar characteristics. Objects belonging to the same entity are bound together in time. In this framework, synchronization between different neurons and desynchronization among different regions perform the binding. In the present work, we implement the temporal correlation to bind auditory image objects. The binding merges the segmented auditory objects belonging to the same source.

3. PROPOSED SYSTEM STRATEGY

Two representations are simultaneously generated: amplitude modulation map, which we call Cochleotopic/AMtopic (CAM) Map[4] and the Cochleotopic/Spectrotopic Map (CSM) that encodes the averaged spectral energies of the cochlear filter bank output. The first representation somewhat reproduces the AM processing performed by multipolar cells (Chopper-S) from the anteroventral cochlear nucleus [19], while the second representation could be closer to the spherical bushy cell processing from the ventral cochlear nucleus areas [18].

We assume that different sources are disjoint in the auditory image representation space and that masking (binary gain) of the undesired sources is feasible. Speech has a specific structure that is different from that of most noises and perturbations [26]. Also, when dealing with simultaneous speakers, separation is possible when preserving the time structure (the probability at a given instant t to observe overlap in pitch and timbre is relatively low). Therefore, a binary gain can be used to suppress the interference (or separate all sources with adaptive masks).

4. DETAILED DESCRIPTION

4.1. Signal analysis

Our CAM/CSM generation algorithm is as follows.

(1) Down-sample to 8000 samples/s.
(2) Filter the sound source using a 256-filter Bark-scaled cochlear filter bank ranging from 100 Hz to 3.6 kHz.
(3) (i) For CAM, extract the envelope (AM demodulation) for channels 30–256; for other low-frequency channels (1–29) use raw outputs.[5]
    (ii) For CSM, nothing is done in this step.
(4) Compute the STFT of the envelopes (CAM) or of the filter bank outputs (CSM) using a Hamming window.[6]
(5) To increase the spectro-temporal resolution of the STFT, find the reassigned spectrum of the STFT [28] (this consists of applying an affine transform to the points to realocate the spectrum).
(6) 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.[7]

[3] Other features like transients, on-, off-responses are observed, but are not implemented here.

[4] To some extent, it is related to modulation spectrograms. See for example work in [24, 25].

[5] Low-frequency channels are said to resolve the harmonics while others do not, suggesting a different strategy for low-frequency channels [27].

[6] Nonoverlapping adjacent windows with 4-millisecond or 32-millisecond length have been tested.

[7] $\log(e_1 + e_2) \simeq \max(\log e_1, \log e_2)$ (unless e_1 and e_2 are both large and almost equal) [4].

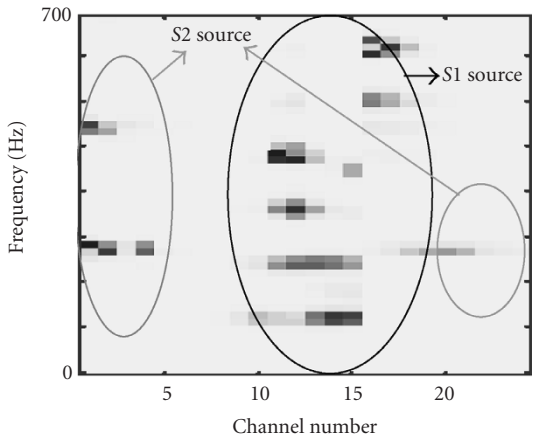

Figure 3: Example of a 24-channel CAM for a mixture of /di/ and /da/ pronounced by two speakers; mixture at SNR = 0 dB and frame center at $t = 166$ milliseconds.

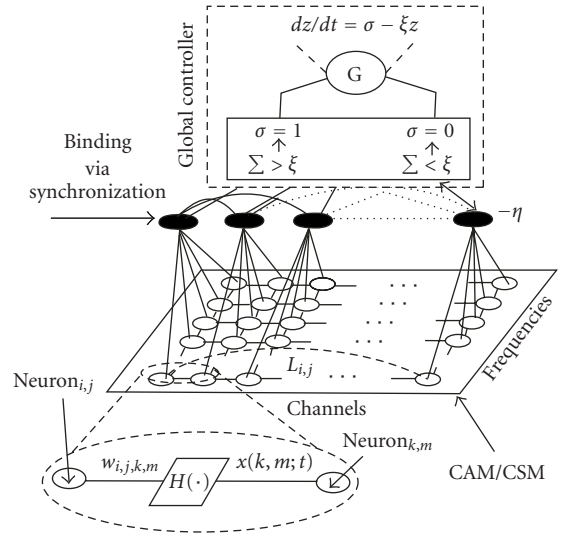

Figure 4: Architecture of the two-layer bio-inspired neural network. G stands for global controller (the global controller for the first layer is not shown on the figure). One long-range connection is shown. Parameters of the controller and of the input layer are also illustrated in the zoomed areas.

It is 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 [29]. To a certain extent, one can imagine that envelope detection and selection between the CAM and the CSM, in the auditory pathway, could be associated to the efferent system in combination with cochlear nucleus processing [30, 31]. For now, in the present experimental setup, selection between the two auditory images is done manually. Figure 3 is an example of a CAM computed through a 24-cochlear-channel filter bank for a /di/ and /da/ mixture pronounced by a female and male speaker. Ellipses outline the auditory objects.

4.2. The neural network

4.2.1. First layer: image segmentation

The dynamics of the neurons we use is governed by a modified version of the Van der Pol relaxation oscillator (Wang-Terman oscillators [7]). The state-space equations for these dynamics are as follows:

$$\frac{dx}{dt} = 3x - x^3 + 2 - y + \rho + p + S, \tag{1}$$

$$\frac{dy}{dt} = \epsilon\left[\gamma\left(1 + \tanh\left(\frac{x}{\beta}\right)\right) - y\right], \tag{2}$$

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.[8] The Euler integration method is used to solve the equations. The first layer is a partially connected network of relaxation oscillators [7]. 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 256 points computed for the FFT are down-sampled to 50 points. Therefore, the first layer consists of 256×50 neurons. The geometric interpretation of pitch (ray distance criterion) is less clear for the first 29 channels, where harmonics are usually resolved.[9] 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.

The weight, $w_{i,j,k,m}(t)$ (Figure 4), between neuron(i, j) and neuron(k, m) of the first layer is

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

where $p(i, j)$ 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

[8] In our simulation, $\epsilon = 0.02$, $\gamma = 4$, $\beta = 0.1$, and $\rho = 0.02$.

[9] Envelopes of resolved harmonics are nearly constants.

is used for *long-range clear-to-confusion zone* connections (6) in CAM processing case. The coupling $S_{i,j}$ defined in (1) is

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

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

$$G(t) = \alpha H(z-\theta), \quad \frac{dz}{dt} = \sigma - \xi z, \tag{5}$$

where σ is equal to 1 if the global activity of the network is greater than a predefined ζ and is zero otherwise (Figure 4). α and ξ are constants.[10]

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

$$L_{i,j}(t) = \begin{cases} 0, & j \geq 30, \\ \sum_{k=225\cdots256} w_{i,j,i,k}(t)H(x(i,k;t)), & j < 30. \end{cases} \tag{6}$$

κ is a binary variable defined as follows:

$$\kappa = \begin{cases} 1 & \text{for CAM}, \\ 0 & \text{for CSM}. \end{cases} \tag{7}$$

4.2.2. Second layer: temporal correlation and multiplicative synapses

The second layer is an array of 256 neurons (one for each channel). Each neuron receives the weighted product of the outputs of the first layer neurons along the frequency axis of the CAM/CSM. The weights between layer one and layer two are defined as $w_{ll}(i) = \alpha/i$, where i can be related to the frequency bins of the STFT and α is a constant for the CAM case, since we are looking for structured patterns. For the CSM, $w_{ll}(i) = \alpha$ is constant along the frequency bins as we are looking for energy bursts.[11] Therefore, the input stimulus to neuron(j) in the second layer is defined as follows:

$$\theta(j;t) = \overline{\prod_i w_{ll}(i)\Xi\{x(i,j;t)\}}. \tag{8}$$

The operator Ξ is defined as

$$\Xi\{x(i,j;t)\} = \begin{cases} 1 & \text{for } x(i,j;t) = 0, \\ x(i,j;t) & \text{elsewhere}, \end{cases} \tag{9}$$

where $\overline{(\cdot)}$ is the *averaging over a time window* operator (the duration of the window is in the order of the discharge period). The multiplication is done only for nonzero outputs (in which spike is present) [32, 33]. This behavior has been observed in the integration of ITD (interaural time difference) and ILD (interlevel difference) information in the barn owl's auditory system [32] or in the monkey's posterior parietal lobe neurons that show *receptive fields* that can be explained by a multiplication of retinal and eye or head position signals [34].

The synaptic weights inside the second layer are adjusted through the following rule:

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

where 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 dynamics of the second layer is given by an equation similar to (4) (without long-range coupling). 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 [7].

The selection strategy at the output of the second layer is based on temporal correlation: neurons belonging to the same source synchronize (same spiking phase) and neurons belonging to other sources desynchronize (different spiking phase).

4.3. Masking and synthesis

Time-reversed outputs of the *analysis* filter bank are passed through the *synthesis* filter bank giving birth to $z_i(t)$. Based on the phase synchronization described in the previous section, a mask is generated by associating zeros and ones to different channels:

$$s(t) = \sum_{i=1}^{256} m_i(t)z_i(t), \tag{11}$$

where $s(N-t)$ is the recovered signal (N is the length of the signal in discrete mode), $z_i(t)$ is the synthesis filter bank output for channel i, and $m_i(t)$ is the mask value. Energy is normalized in order to have same SPL for all frames. Note that two-source mixtures are considered throughout this article but the technique can be potentially used for more sources. In that case, for each time frame n, labeling of individual channels is equivalent to the use of multiple masks (one for each source).

5. EXPERIMENTS

We first illustrate the separation of two simultaneous speakers (double-voiced speech segregation), separation of a speech sentence from an interfering siren, and then compare with other approaches.

The magnitude of the CAM's STFT is a structured image whose characteristics depend heavily on pitch and formants. Therefore, in that representation, harmonic signals are separable. On the other hand, the CSM representation is more suitable for inharmonic signals with bursts of energy.

[10] $\zeta = 0.2$, $\alpha = -0.1$, $\xi = 0.4$, $\eta = 0.05$, and $\theta = 0.9$.

[11] In our simulation, $\alpha = 1$.

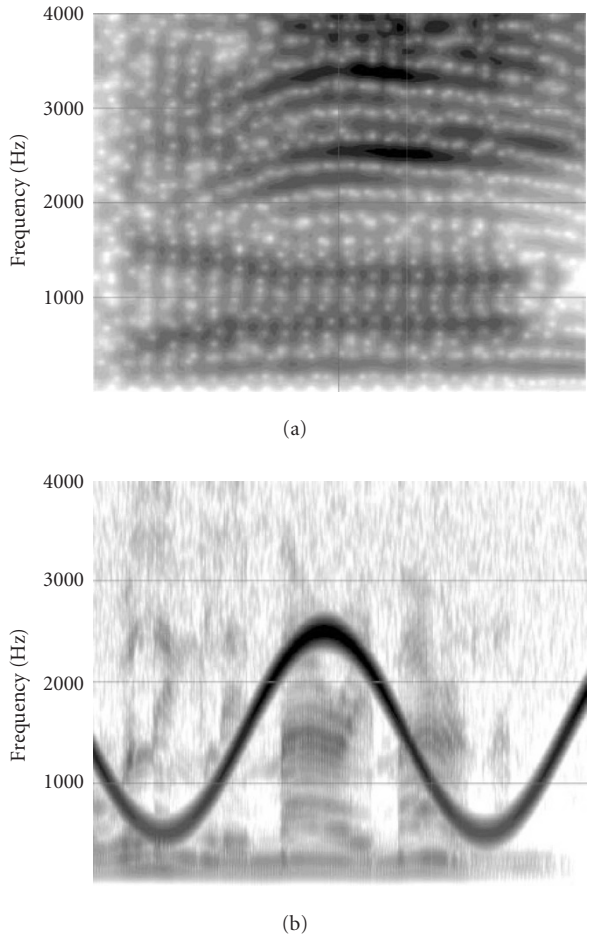

Figure 5: (a) Spectrogram of the /di/ and /da/ mixture. (b) Spectrogram of the sentence "I willingly marry Marilyn" plus siren mixture.

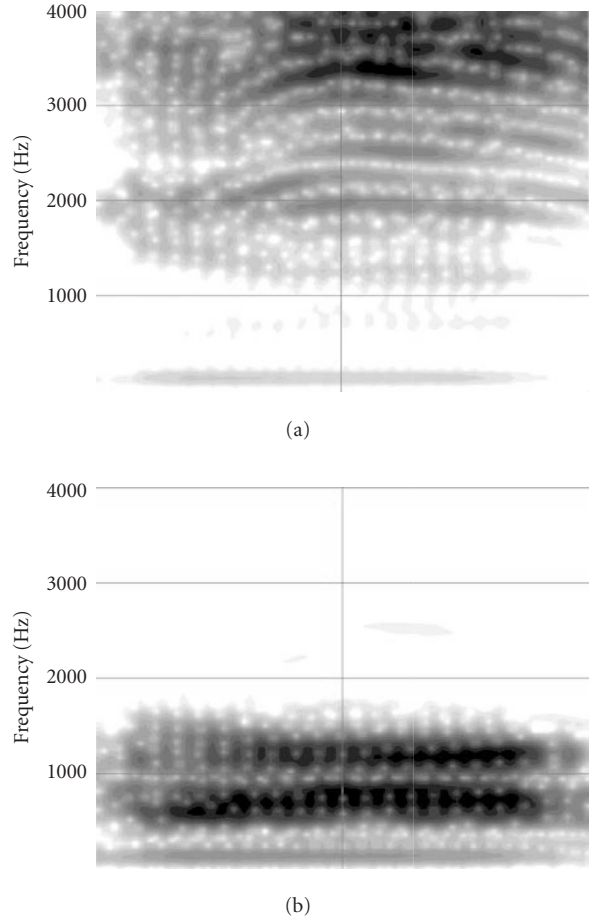

Figure 6: (a) The spectrogram of the extracted /di/. (b) The spectrogram of the extracted /da/.

5.1. Double-speech segregation case

Two speakers have simultaneously and respectively pronounced a /di/ and a /da/ (spectrogram Figure 5a). We observed that the CSM representation does not generate very discriminative representation while, from the CAM, the 2 speakers are well separable (see Figure 6). After binding, two sets of synchronized neurons are obtained: one for each speaker. Separation is performed by using (11), where $m_i(t) = 0$ for one speaker and $m_i(t) = 1$ for the other speaker (target speaker).

5.2. Sentence plus siren

A modified version of the siren used in Cooke's database [7] (http://www.dcs.shef.ac.uk/~martin/) is mixed with the sentence "I willingly marry Marilyn." The spectrogram of the mixed sound is shown in Figure 5b.

In that situation, we look at short but high energy bursts. The CSM representation generates a very discriminative representation of the speech and siren signals, while, on the other hand, the CAM fades the image as the envelopes of the interfering siren are not highly modulated. After binding, two sets of synchronized neurons are obtained: one for each source. Separation is performed by using (11), where $m_i(t) = 0$ for the siren and $m_i(t) = 1$ for the speech sentence and vice versa.

5.3. Comparisons

Three approaches are used for comparison: the methods proposed by Wang and Brown [7] (W-B), by Hu and Wang [8] (H-W), and by Jang and Lee [35] (J-L). W-B uses an oscillatory neural network but relies on pitch information through correlation, H-W uses a multipitch tracking system, and J-L needs statistical estimation to perform the MAP-based separation.

6. RESULTS

Results can be heard and evaluated at http://www-edu.gel.usherbrooke.ca/pichevar/, http://www.gel.usherb.ca/rouat/.

6.1. Siren plus sentence

The CSM is presented to the spiking neural network. The weighted product of the outputs of the first layer along the

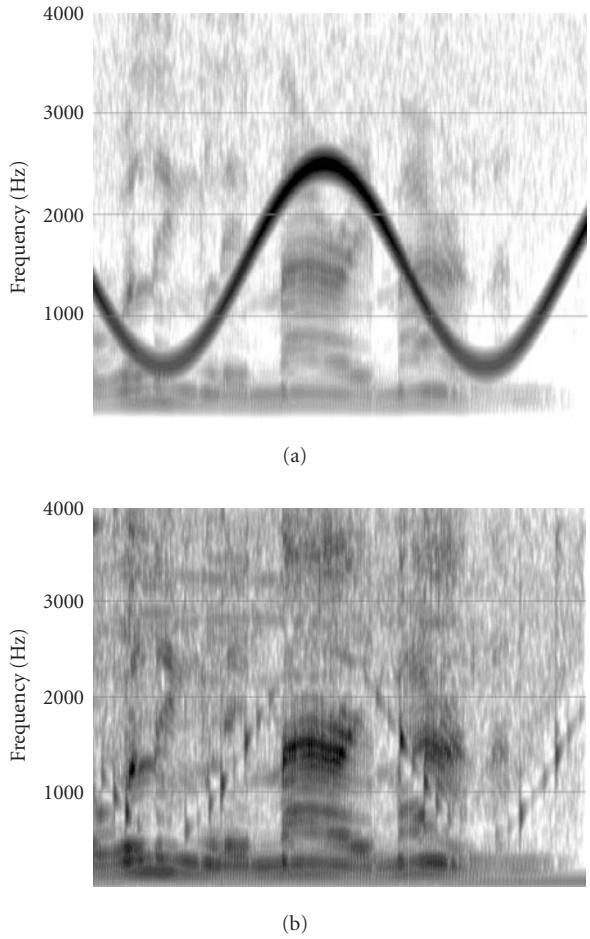

Figure 7: (a) The spectrogram of the extracted siren. (b) The spectrogram of the extracted utterance.

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 Figure 7. A CSM is extracted every 10 milliseconds and the selection is made by 10-millisecond intervals. In a future work, we will use much smaller selection intervals and shorter STFT windows to prevent discontinuities, as observed in Figure 7.

6.2. Double-voiced speech

Perceptual tests have shown that although we reduce sound quality after the process, the vowels are separated and are clearly recognizable.

6.3. Evaluation and comparisons

Table 1 reports the perceptive evaluation of speech quality criterion (PESQ) on sentences corrupted with various noises. The first column is the intruding noise, the second column gives the initial SNR of the mixtures, and other columns are the PESQ scores for the reference methods. Table 2 gives the comparison for a female speech sentence corrupted with rock music (http://home.bawi.org/~jangbal/research/demos/rbss1/sepres.html).

Table 1: PESQ for three different methods: P-R (our proposed approach), W-B [7], and H-W [8]. The intrusion noises are (a) 1 kHz pure tone, (b) FM siren, (c) telephone ring, (d) white noise, (e) male-speaker intrusion (/di/) for the French /di//da/ mixture, and (f) female-speaker intrusion (/da/) for the French /di//da/ mixture. Except for the last two tests, the intrusions are mixed with a sentence taken from Martin Cooke's database.

Intrusion (noise)	Ini. SNR mixture	P-R (PESQ)	W-B (PESQ)	H-W (PESQ)
Tone	−2 dB	0.403	0.223	0.361
Siren	−5 dB	2.140	1.640	1.240
Telephone ring	3 dB	0.860	0.700	0.900
White	−5 dB	0.880	0.223	0.336
Male (da)	0 dB	2.089	N/A	N/A
Female (di)	0 dB	0.723	N/A	N/A

Table 2: PESQ for two different methods: P-R (our proposed approach) and J-L [35]. The mixture comprises a female voice with musical background (rock music).

Mixture	Separated sources	P-R (PESQ)	J-L (PESQ)
Music & female (AF)	Music	1.724	0.346
	Voice	0.550	0.630

Many criteria are used in the literature to compare sound source separation performance. Some of the most important are SNR, segmental SNR, PEL (percentage of energy loss), PNR (percentage of noise residue), and LSD (log-spectral distortion). As they do not take into account perception, we propose to use another criterion, that is, the PESQ, to better reflect human perception. The PESQ (perceptual evaluation of speech quality) is an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. The key to this process is the transformation of both the original and degraded signals into an internal representation that is similar to the psychophysical representation of audio signals in the human auditory system, taking into account the perceptual frequency (Bark scale) and loudness (sone). This allows a small number of quality indicators to be used to model all subjective effects. These perceptual parameters are combined to create an objective listening quality MOS. The final score is given on a range of −0.5 to 4.5.[12]

In all cases, the system performs better than W-P [7] and H-W [8], except for the telephone ring intrusion where H-W is slightly better. For the double-voiced speech, the male speaker is relatively well extracted. Other evaluations we made are based on LSD and SNR and also converge to similar results.

[12] 0 corresponds to the worst quality and 4.5 corresponds to the best quality (no degradation).

7. CONCLUSION AND FURTHER WORK

Based on evidences regarding the dynamics of the efferent loops and on the richness of the representations observed in the cochlear nucleus, we proposed a technique to explore the monophonic source separation problem using a multirepresentation (CAM/CSM) bio-inspired preprocessing stage and a bio-inspired neural network that does not require any a priori knowledge of the signal.

For the time being, the CSM/CAM selection is made manually. In a near future, we will include a top-down module based on the local SNR gain to selectively find the suitable auditory image representation, also depending on the neural network synchronization.

In the reported experiments, we segregate two sources to illustrate the work, but the approach is not restricted to that number of sources.

Results obtained from signal synthesis are encouraging and we believe that spiking neural networks in combination with suitable signal representations have a strong potential in speech and audio processing. The evaluation scores show that our system yields fairly comparable (and most of the time better) performance than other methods even if it does not need a priori knowledge and is not limited to harmonic signals.

ACKNOWLEDGMENTS

This work has been funded by NSERC, MRST of Québec Government, Université de Sherbrooke, and by Université du Québec à Chicoutimi. Many thanks to DeLiang Wang for fruitful discussions on oscillatory neurons, to Wolfgang Maass for pointing the work by Milner, to Christian Giguère for discussions on auditory pathways, and to the anonymous reviewers for constructive comments.

REFERENCES

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

[2] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.

[3] S. T. Roweis, "One microphone source seperation," in *Proc. Neural Information Processing Systems (NIPS '00)*, pp. 793–799, Denver, Colo, USA, 2000.

[4] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 1009–1012, Geneva, Switzerland, September 2003.

[5] M. J. Reyes-Gomez, B. Raj, and D. R. W. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 1, pp. 664–667, Hong Kong, China, April 2003.

[6] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

[7] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[8] G. Hu and D. Wang, "Separation of stop consonants," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 2, pp. 749–752, Hong Kong, China, April 2003.

[9] P. Milner, "A model for visual shape recognition," *Psychological Review*, vol. 81, no. 6, pp. 521–535, 1974.

[10] C. von der Malsburg, "The correlation theory of brain function," Internal. Rep. 81-2, Max-Planck Institute for Biophysical Chemistry, Gottingen, Germany, 1981.

[11] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[12] D. E. Haines, Ed., *Fondamental Neuroscience*, Churchill Livingstone, San Diego, Calif, USA, 1997.

[13] C. E. Schreiner and J. V. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory filed (AAF)," *Hearing Research*, vol. 21, no. 3, pp. 227–241, 1986.

[14] C. Schreiner and G. Langner, "Periodicity coding in the inferior colliculus of the cat. II. Topographical organization," *Journal of Neurophysiology*, vol. 60, no. 6, pp. 1823–1840, 1988.

[15] L. Robles, M. A. Ruggero, and N. C. Rich, "Two-tone distortion in the basilar membrane of the cochlea," *Nature*, vol. 349, pp. 413–414, 1991.

[16] E. F. Evans, "Auditory processing of complex sounds: an overview," in *Phil. Trans. Royal Society of London*, pp. 1–12, Oxford Press, Oxford, UK, 1992.

[17] M. A. Ruggero, L. Robles, N. C. Rich, and A. Recio, "Basilar membrane responses to two-tone and broadband stimuli," in *Phil. Trans. Royal Society of London*, pp. 13–21, Oxford Press, Oxford, UK, 1992.

[18] C. K. Henkel, "The auditory system," in *Fondamental Neuroscience*, D. E. Haines, Ed., Churchill Livingstone, New York, NY, USA, 1997.

[19] P. Tang and J. Rouat, "Modeling neurons in the anteroventral cochlear nucleus for amplitude modulation (AM) processing: application to speech sound," in *Proc. 4th IEEE International Conf. on Spoken Language Processing (ICSLP '96)*, vol. 1, pp. 562–565, Philadelphia, Pa , USA, October 1996.

[20] A. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, Mass, USA, 1994.

[21] M. W. Beauvois and R. Meddis, "A computer model of auditory stream segregation," *The Quaterly Journal of Experimental Psychology*, vol. 43, no. 3, pp. 517–541, 1991.

[22] C. von der Malsburg and W. Schneider, "A neural cocktail-party processor," *Biological Cybernetics*, vol. 54, pp. 29–40, 1986.

[23] C. von der Malsburg, "The what and why of binding: the modeler's perspective," *Neuron*, vol. 24, no. 1, pp. 95–104, 1999.

[24] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[25] G. Meyer, D. Yang, and W. Ainsworth, "Applying a model of concurrent vowel segregation to real speech," in *Computational Models of Auditory Function*, S. Greenberg and M. Slaney, Eds., pp. 297–310, IOS Press, Amsterdam, The Netherlands, 2001.

[26] J. Rouat, "Spatio-temporal pattern recognition with neural networks: application to speech," in *Proc. International Conference on Artificial Neural Networks (ICANN '97)*, vol 1327 of *Lecture Notes in Computer Science*, pp. 43–48, Springer, Lausanne, Switzerland, October 1997.
[27] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, 1997.
[28] F. Plante, G. Meyer, and W. A. Ainsworth, "Improvement of speech spectrogram accuracy by the method of reassignment," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 282–287, 1998.
[29] 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*, vol. 7, pp. 348–357, 2002.
[30] C. Giguere and P. C. Woodland, "A computational model of the auditory periphery for speech and hearing research," *Journal of the Acoustical Society of America*, vol. 95, pp. 331–349, 1994.
[31] M. Liberman, S. Puria, and J. J. Guinan, "The ipsilaterally evoked olivocochlearreflex causes rapid adaptation of the 2f1-f2 distortion product otoacoustic emission," *Journal of the Acoustical Society of America*, vol. 99, pp. 2572–3584, 1996.
[32] F. Gabbiani, H. Krapp, C. Koch, and G. Laurent, "Multiplicative computation in a visual neuron sensitive to looming," *Nature*, vol. 420, pp. 320–324, 2002.
[33] J. Pena and M. Konishi, "Auditory spatial receptive fields created by multiplication," *Science*, vol. 292, pp. 294–252, 2001.
[34] R. Andersen, L. Snyder, D. Bradley, and J. Xing, "Multimodal representation of space in the posterior parietal cortex and its use in planning movements," *Annual Review of Neuroscience*, vol. 20, pp. 303–330, 1997.
[35] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Single-channel signal separation using time-domain basis functions," *IEEE Signal Processing Letters*, vol. 10, no. 6, pp. 168–171, 2003.

Jean Rouat holds an M.S. degree in physics from University de Bretagne, France (1981), an E. & E. M.S. degree in speech coding and speech recognition from Université de Sherbrooke (1984), and an E. & E. Ph.D. degree in cognitive and statistical speech recognition jointly from Université de Sherbrooke and McGill University (1988). From 1988 to 2001 he was with the Université du Québec à Chicoutimi (UQAC). In 1995 and 1996, he was on a sabbatical leave with the Medical Research Council, Applied Psychological Unit, Cambridge, UK, and the Institute of Physiology, Lausanne, Switzerland. In 1990 he founded the ERMETIS, Microelectronics and Signal Processing Research Group, UQAC. He is now with Université de Sherbrooke where he founded the Computational Neuroscience and Signal Processing Research Group. He regularly acts as a reviewer for speech, neural networks, and signal processing journals. He is an active member of scientific associations (Acoustical Society of America, International Speech Communication, IEEE, International Neural Networks Society, Association for Research in Otolaryngology, ACM, etc.). He is a Member of the IEEE Technical Committee on Machine Learning for Signal Processing.

Ramin Pichevar was born in March 1974, in Paris, France. He received his B.S. degree in electrical engineering (electronics) in 1996 and the M.S. degree in electrical engineering (telecommunication systems) in 1999, both in Tehran, Iran. He received his Ph.D. degree in electrical and computer engineering from Université de Sherbrooke, Québec, Canada, in 2004. During his Ph.D., he gave courses on signal processing and computer hardware as a Lecturer. In 2001 and 2002 he did two summer internships at Ohio State University, USA, and at the University of Grenoble, France, respectively. He is now a Postdoctoral Fellow and Research Associate in the Computational Neuroscience and Signal Processing Laboratory at the University of Sherbrooke under an NSERC (National Sciences and Engineering Council of Canada) Ideas to Innovation (I2I) grant. His domains of interest are signal processing, computational auditory scene analysis (CASA), neural networks with emphasis on bio-inspired neurons, speech recognition, digital communications, discrete-event simulation, and image processing.