Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2007) 109–120
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Monophonic sound source separation with an unsupervised network of spiking neurones

Ramin Pichevar*,[1], Jean Rouat

Département de génie électrique et génie informatique, Université de Sherbrooke, 2500 boul. de l'Université, Sherbrooke, Qué., Canada J1K 2R1

Available online 8 August 2007

Abstract

We incorporate auditory-based features into an unconventional pattern classification system, consisting of a network of spiking neurones with dynamical and multiplicative synapses. Although the network does not need any training and is autonomous, the analysis is dynamic and capable of extracting multiple features and maps. The neural network allows computing a binary mask that acts as a dynamic switch on a speech vocoder made of an FIR gammatone analysis/synthesis bank of 256 filters. We report experiments on separation of speech from various intruding sounds (siren, telephone bell, speech, etc.) and compare our approach to other techniques by using the log spectral distortion (LSD) metric.

Keywords: Amplitude modulation; Auditory scene analysis; Auditory maps; Source separation; Speech enhancement; Spikes; Neurones

1. Introduction

Separation of mixed signals is a major issue in many applications in the context of audio processing. It can be used, for instance, to assist a robot in segregating multiple speakers, to ease the automatic transcription of video via the audio tracks, to separate musical instruments before automatic transcription; to clean a signal before performing speech recognition on it, etc. The ideal instrumental setup would use an array of microphones during the recording phase in order to gather many audio channels. Unfortunately, in many situations, only one channel is available to the audio engineer that has to solve the separation problem. As such, the automatic separation of the sources is much more difficult. Most of the monophonic systems described in the scientific literature perform reasonably well on specific signals (generally voiced speech), but fail to efficiently separate a broad range of signals. These relatively unsatisfactory results may be enhanced by exploiting expertise and knowledge of engineering, psychoacoustics, physiology, and computer science.

In this paper we assume that monophonic source separation systems consist of two main stages: auditory map generation and auditory grouping. Our approach is based on auditory scene analysis. An auditory map is a visual representation of how a sound mixture is perceived in the brain.

2. Auditory scene analysis

A remarkable feat of the auditory system is its ability to disentangle the acoustic mixture and group the acoustic energy from the same event. This fundamental process of auditory perception is called auditory scene analysis. Of particular importance in auditory scene analysis is the separation of speech from interfering sounds, or speech segregation.

2.1. Auditory scene analysis according to Bregman

Bregman [7] defines the concept of "auditory streams", i.e., the mental percept of a succession of auditory events.

*Corresponding author. Tel.: +1 819 821 8000x2187; fax: +1 819 821 7937.
E-mail addresses: Ramin.Pichevar@usherbrooke.ca (R. Pichevar), Jean.Rouat@usherbrooke.ca (J. Rouat).
URLs: http://www-edu.gel.usherb.ca/picr1601 (R. Pichevar), http://www.gel.usherb.ca/rouat (J. Rouat).
[1]Now with the Communications Research Centre, Ottawa, Canada.

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.08.001

Auditory streaming entails two complementary domains of study. The first one tries to determine how sounds cohere to create a sense of continuation. That is the subject of stream fusion. Since more than one source can sound concurrently, the second domain of study examines how concurrent activities retain their independent identities; that is the subject of stream segregation. For instance, auditory streaming is important in assigning consecutive speech elements to the same speaker, or in following a melodic line in a background of other musical sounds.

In this paper, auditory stream segregation and integration are performed by our proposed neural network.

2.2. Computational auditory scene analysis

Computational auditory scene analysis (CASA) is an attempt to realise auditory scene analysis systems with computers. In Bregman's terminology, bottom-up processing corresponds to primitive processing, and top-down processing means schema-based processing. The auditory cues proposed by Bregman for simple tones are not directly applicable to complex sounds. Therefore, one should develop more sophisticated cues based on different auditory maps.

One of the first attempts to perform CASA has been done by Weintraub [65]. Ellis [13] uses sinusoidal tracks created by the interpolation of the spectral peaks of the output of a cochlear filterbank. Mellinger's model [37] uses partials. A partial is formed if an activity on the onset maps (the beginning of an energy burst) coincides with an energy local minimum of the spectral maps. Using the aforementioned assumption Mellinger proposed a CASA system in order to separate musical instruments. Cooke [9] has introduced the "synchrony strands", which is the counterpart of Mellinger's cues in speech. The integration and segregation of streams is done using Gestalt and Bregman's heuristics. Berthommier and Meyer use amplitude modulation maps [5] (see also [60,44,38]). Gaillard [15] follows a more conventional approach by using the first zero crossings for the detection of pitch and harmonic structures in the frequency–time map. Brown's algorithm [8] is based on the mutual exclusivity Gestalt principle. Hu and Wang use a pitch tracking technique [23]. Wang and Brown [63] use correlograms in combination with bio-inspired neural networks. Grossberg et al. [18] propose a neural architecture that implements Bregman's rules for simple sounds. Sameti et al. [57] use hidden Markov models (HMM), while in [56,55,48] factorial HMMs are used. Jang and Lee [28] use a technique based on maximum a posteriori (MAP) criterion. For another probability-based CASA see [12].

Irino and Patterson [25] propose an auditory representation that is synchronous to the glottis and preserves fine temporal information. Their representation makes the synchronous segregation of speech possible. In [19] a model of multi-resolution with both high- and low-resolution representations of the audio signal in parallel is used. The authors propose an implementation for speech recognition. Nix et al. [40] perform a binaural statistical estimation of two speech sources by an approach that integrates temporal- and frequency-specific features of speech. It tracks magnitude spectra and direction on a frame-by-frame basis.

Most of the aforementioned systems require training and are supervised.[2] Other works are reported in [51,11].

3. Segregation and integration with binding

Neurone assemblies (groups) of spiking neurones can be used to implement segregation and fusion (integration) of objects into an auditory image representation. Usually, in signal processing, correlations (or distances) between signals are implemented with delay lines, products and summations. With spiking neurones, comparison (temporal correlation) between signals can be made without the implementation of delay lines. This has been achieved by presenting auditory images to spiking neurones with dynamic synapses. Then, a spontaneous organisation appears in the network by a set of neurones firing in synchrony. Neurones with the same firing phase belong to the same auditory object. In 1976 and 1981, the temporal correlation that performs binding was proposed by Milner in [39] and by Malsburg [34–36]. They observed that synchrony is a crucial feature to bind neurones associated to similar characteristics. Objects belonging to the same entity are bound together in time. In other words, synchronisation between different neurones and desynchronisation among different regions perform the binding. On the other hand, there are other approaches such as the hierarchical coding, or attentional models that try to find a solution without using the synchronisation concept (for a review see [49]). To a certain extent, such property has been exploited in [6] to perform unsupervised clustering for recognition on images, in [58] for vowel processing with spike synchrony between cochlear channels, in [21] to propose pattern recognition with spiking neurones, and in [32] to perform cell assembly of spiking neurones using Hebbian learning with depression. In [64] an efficient and robust technique for image segmentation is presented. Wang and Brown [63] studied the potential application of neural synchrony in CASA.

Bio-inspired neural networks are well adapted to signal processing whenever processing time is an important issue. They do not always require a training period and can work in a fully unsupervised mode. Adaptive and unsupervised recognition of sequences is a crucial property of living neurones. Among the many properties listed in this section, this paper implements the segregation and integration with sets of synchronous neurones.

[2] This is not the case for the proposed neural network architecture.

4. Proposed system strategy

We propose a bio-inspired bottom-up CASA system. Fig. 1 shows the building blocks of this approach. The sound mixture is filtered by an FIR gammatone filterbank giving birth to 256 different signals, each belonging to one of the cochlear channels. We propose two different representations: the CAM (Cochleotopic/AMtopic Map) and the CSM (Cochleotopic/Spectropic Map) as described in Section 6. Depending on the nature of the intruding sound (speech, music, noise, etc.) one of the maps is selected as explained later in Section 6. The map is applied to our proposed two-layered spiking neural network (Section 7). We propose to generate a binary mask based on the neural synchrony in the output of the neural network (Section 8). The binary mask is then multiplied with the output of the FIR gammatone synthesis filterbank and the channels are summed up.

Our bio-inspired approach has the following advantages, when compared to other non-bio-inspired techniques:

- It does not need the knowledge of explicit rules to create the separation mask. In fact, finding a suitable extension of the rules developed by Bregman for simple sounds to the real world is difficult. Therefore, as long as the aforementioned rules are not derived and well documented [11], expert systems will be difficult to use.
- It does not need any time-consuming training phase prior to the separation phase. This is contrary to approaches based on statistics like HMMs [57], factorial HMMs [48,56] or MAP [28].
- It is autonomous as it does not use hierarchical classification [35].

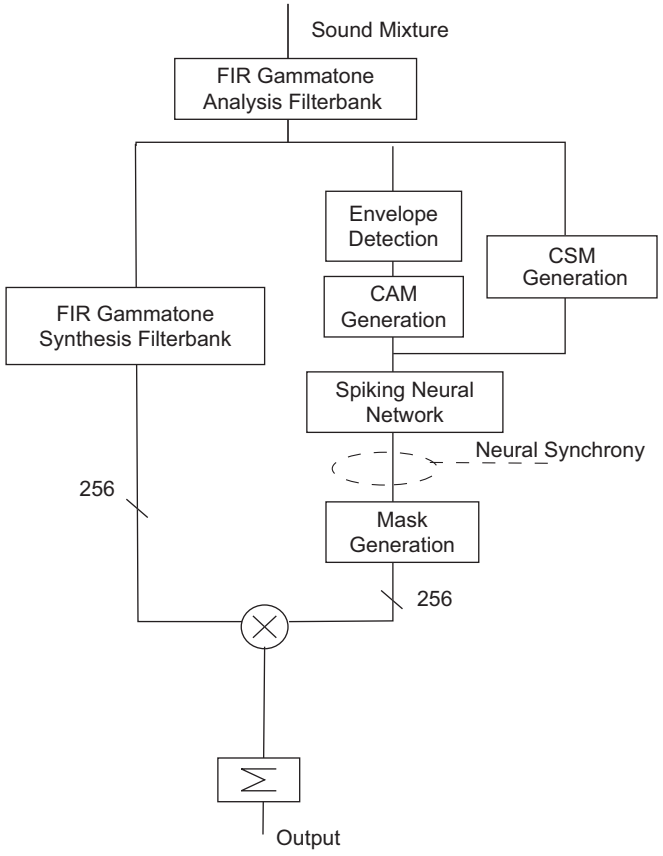

Fig. 1. The block diagram for the proposed bio-inspired sound source separation technique.

The work by Wang and Brown [63] is a breakthrough in the field of bio-inspired neural network CASA and, to our knowledge, is one of the first in the field. Our work uses the same oscillatory neurones as in [63], but with a different neural architecture and a different preprocessing stage. We propose two-dimensional representation maps that do not require pitch estimation. Moreover, the signal is continuously presented to the system, i.e., no segmentation is required. Our proposed architecture uses a 2-layered neural network, in which the second layer is one-dimensional. We use an FIR gammatone filterbank with much less synthesis distortion [46].

5. Analysis/synthesis filterbank

We use an FIR implementation of the well-known gammatone filterbank [41] as the analysis/synthesis filterbank. Two hundred and fifty six channels having centre frequencies from 100 to 3600 Hz and uniformly spaced on an equivalent rectangular bandwidth (ERB) scale [41] are used. The sampling rate is 8000 samples/s.

The actual time-varying filtering is done by the mask. Our technique is based on generating masks obtained by grouping synchronous oscillators of the neural net (see Section 7.4). The outputs of the synthesis filterbank are multiplied by the mask (defined in Section 7). Thus, auditory channels belonging to interfering sound sources are muted and channels belonging to the sound source of interest remain unaffected.[3]

Before the signals of the masked auditory channels are added to form the synthesised signal, they are passed through the synthesis filters, whose impulse responses are time-reversed versions of the impulse responses of the corresponding analysis filters.[4]

This non-decimated FIR analysis/synthesis filterbank was proposed in [26] and also used in a perceptual speech coder in [31] (in the latter only 20 channels were used). For more details on the design of the analysis/synthesis filterbank see [46].

6. Analysis and auditory image generation

We propose two spectral two-dimensional representations that are generated simultaneously:

[3]This is, equivalent to labelling—for each time frame—cochlear channels. A value of one is associated to the target signal and a value of 0 to the interfering signal.

[4]For cancelling out cochlear channel delays.

- A modified version of the well-known and documented amplitude modulation map, which we call CAM Map—closely related to modulation spectrograms as defined by other authors [2,38].
- The CSM that encodes the averaged spectral energies of the cochlear filterbank outputs.

By proposing the CAM, it was desired to somewhat reproduce the AM processing performed by multipolar cells (Chopper-S) from the anteroventral cochlear nucleus [59]. The second representation (CSM) is motivated by the functioning of the spherical bushy cell processing from the ventral cochlear nucleus [20].

Our CAM/CSM generation algorithm is as follows:

(1) Filter the sound source using a 256-filter ERB-scaled cochlear filterbank ranging from 100 to 3.6 kHz.
(2)
   - For CAM: Extract the envelope (AM demodulation) for channels 30–256; for other low frequency channels (1–29) use raw outputs.[5]
   - For CSM: Do nothing in this step.
(3) Compute the STFT of each cochlear output using a Hamming window.[6]
(4) In order to increase the spectro-temporal resolution of the STFT, find the reassigned spectrum of the STFT [47] (this consists of applying an affine transform to the points in order to relocate the spectrum).[7]
(5) Compute the logarithm of the magnitude of the STFT. The logarithm enhances the presence of the stronger source in a given two-dimensional frequency bin of the CAM/CSM.[8]

It has recently been 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]. We suppose here that envelope detection and selection between the CAM and the CSM, in the auditory pathway, could be associated to the change of stiffness of hair cells combined with cochlear nucleus processing [17,33]. For now, in the present experimental setup, selection between the two auditory images is done manually. In near future, we plan to implement efferent feedback control for signal representation adaptation.

Fig. 2 is an example of a CAM computed through a 24 cochlear channels filterbank for a /di/ and /da/ mixture pronounced by a female and male speaker. Ellipses outline the auditory objects.

[5]Low-frequency channels are said to resolve the harmonics while others do not. It suggests a different strategy for low frequency channels [53].

[6]Non-overlapping adjacent windows with 4 or 32 ms lengths have been tested.

[7]Note that one can do a compromise by not computing the time-consuming reassigned spectrum for slightly worse performance.

[8]$\log(e_1+e_2) \simeq \max(\log e_1, \log e_2)$ (unless e_1 and e_2 are both large and almost equal) [55].

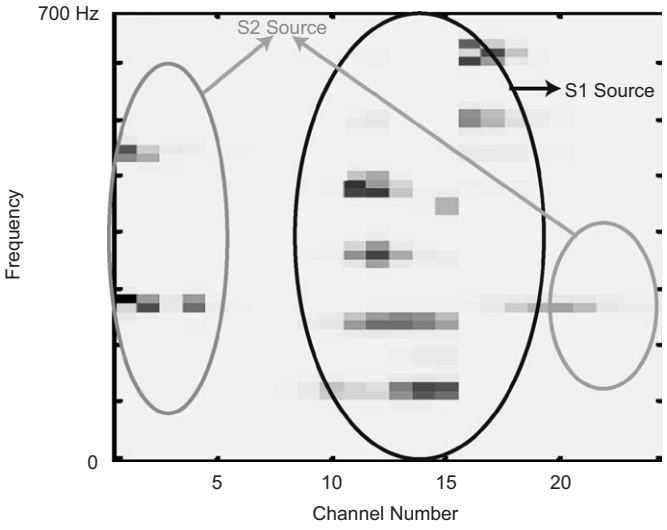

Fig. 2. Example of a 24 channel CAM for a mixture of /di/ and /da/ pronounced by two speakers; mixture at $SNR = 0$ dB and frame centre at $t = 166$ ms. Channels 10–18 belong to one of the sources and other channels belong to the other source.

7. Architecture of the neural network

In this section, we propose a novel neural architecture based on bio-inspired neurones.

7.1. Bio-inspired neural networks

Bio-inspired (Spiking) neural networks try to mimic the behaviour of real neurones in animals and humans. They allow the processing of temporal sequences, contrary to most classical neural networks that are only suitable for static (time-invariant) data.

In the case of bio-inspired neural networks, temporal sequence processing is done naturally because of the intrinsic dynamic behaviour of the neurones. The pioneering work in the field of bio-inspired neural networks has been done by Hodgkin and Huxley at the University of Plymouth [16]. In the 1950s, they came up with a mathematical model to describe the behaviour of a squid axon. Although this model is the most complete so far (it can predict most of the behaviours seen in simple biological neurones), it is very complex and difficult to simulate in an artificial neural network paradigm. Hence, simplified models like the Wang–Terman model described in the following subsection has been proposed in the literature.

7.2. Building blocks of the neural architecture

Building blocks of our architecture are the well-known "Wang–Terman" oscillators [64]. There is an active phase when the neurone spikes and a relaxation phase when the neurone is silent. Information in our network is conveyed in the relative phase of each oscillation (spike).

The state-space equations for this dynamical system are as follows:

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

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

where x is the membrane potential (output) of the neurone 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 neurone, and S is the coupling from other neurones (connections through synaptic weights). ε, γ and β are constants. The Euler integration method is used to solve the equations.

The network consists of two layers. The first layer is two-dimensional and the second layer is one-dimensional.

7.3. First layer: auditory image segmentation

The first layer essentially performs a segmentation of the auditory map image. A good reference on image segmentation with oscillatory neurones can be found in [64]. The dynamics of the neurones we use is governed by a modified version of the Van der Pol relaxation oscillator (Wang–Terman oscillators [63]). The first layer is a partially connected network of relaxation oscillators [63]. Each neurone is connected to its four neighbours. The CAM (or the CSM) is applied to the input of the neurones. Our observations have shown that the geometric interpretation of pitch (ray distance criterion) is less clear for the first 29 channels (channels where harmonics are usually resolved).

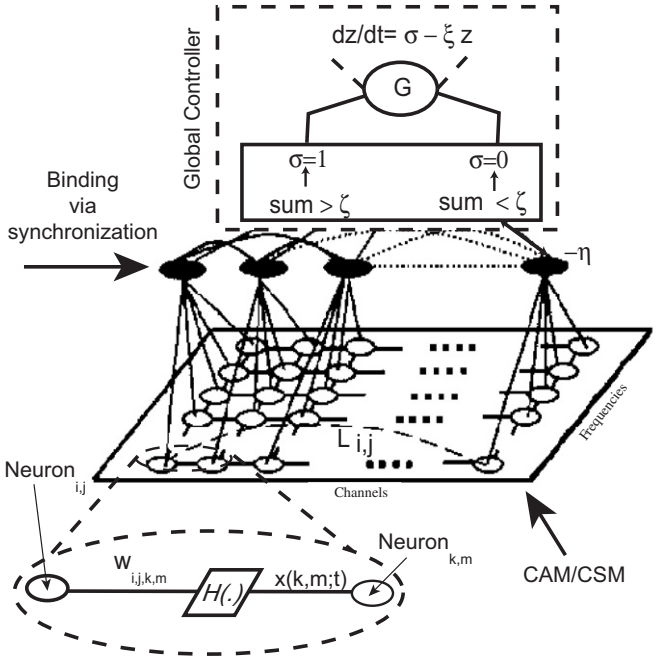

Fig. 3. Architecture of the two-layer bio-inspired neural network. G: Stands for global controller (the global controller for the first layer is not shown in the figure). One long range connection is shown in the figure.

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 helps the network to better extract harmonic patterns (Fig. 3).

The weight at time t between *neurone*(i,j) and *neurone*(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 $p(i,j;t)$ and $p(k,m;t)$ are, respectively, external inputs to *neuron*(i,j) and *neuron*$(k,m) \in N(i,j)$. $Card\{N(i,j)\}$ is a normalisation factor and is equal to the cardinal number (number of elements) of the set $N(i,j)$ containing neighbours connected to the *neurone*(i,j) (can be equal to 4, 3 or 2 depending on the location of the neuron on the map, i.e., centre, corner, etc.). The external input values are normalised. 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. (7)) in CAM processing case. The coupling at time t, $S_{i,j}(t)$ defined in Eq. (1) is

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

$H(.)$ is the Heaviside function and η is a constant. $x(k,m;t)$ is the membrane potential. $G(t)$ is a global controller whose dynamics is governed by the following equation:

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

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

σ is equal to 1 if the global activity of the network is greater than a predefined threshold ζ and is zero otherwise. α, ξ and θ are constants. z is an internal state variable.

Table 1
The numerical values of the different parameters used in the first layer of the network

Constant's name	Value
λ	1
θ	0.9
α	−0.1
ξ	0.4
ζ	0.2
η	0.05
γ	4.0
ε	0.02
ρ	0.02
β	0.1
κ	0.2

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

$$L_{i,j}(t) = \begin{cases} 0, & j \geqslant 30, \\ \sum_{k=225\ldots256} w_{i,j,i,k}(t) H(x(i,k;t)), & j < 30. \end{cases} \quad (7)$$

κ is a binary variable defined as follows:

$$\kappa = \begin{cases} 0.2 & \text{for CAM} \\ 0 & \text{for CSM.} \end{cases} \quad (8)$$

For a given frequency i, channels j that have an index smaller than 30 have connections from channels k indexed between 225 and 256.

The first layer is designed to handle presentations of auditory maps with continuous sliding and overlapping windows on the signal—with real application perspectives in mind (see Table 1 for the parameter values).

7.4. Second layer: temporal correlation and multiplicative synapses

The second layer performs temporal correlation between neurones. Each of the neurones represents a cochlear channel of the analysis/synthesis filterbank. For each presented auditory map, the second layer establishes binding between neurones whose entry is dominated by the same source. The dendrites establish multiplicative synapses with the first layer. The second layer is an array of 256 neurones (one for each channel) similar to those described by Eqs. (1) and (2) in Section 7.3. Each neurone receives the weighted product of the outputs of the first layer neurones 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. Therefore, the input stimulus to *neuron*(j) in the second layer is defined as follows:

$$\theta(j;t) = \overline{\prod_i w_{ll}(i) \Xi\{x(i,j;t)\}}. \quad (9)$$

The operator Ξ is defined as

$$\Xi\{x(i,j;t)\} = \begin{cases} 1 & \text{for } x(i,j;t) = 0, \\ x(i,j;t) & \text{elsewhere,} \end{cases} \quad (10)$$

where $\overline{()}$ is the *averaging over a time window* operator (the duration of the window is on the order of the discharge period). The multiplication with $x(i,j;t)$ is done only for non-zero $x(i,j;t)$ (outputs of the first layer) (in which a spike is present) [14,42]. This behaviour has been observed in the integration of interaural time difference (ITD) and inter level difference (ILD) information in the barn owl's auditory system [14] or in the monkey's posterior parietal lobe neurones that show *receptive fields* that can be explained by a multiplication of retinal and eye or head position signals [1].

Table 2
The numerical values of the different parameters used in the second layer of the network

Constant's name	Value
α	1
μ	2

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

$$w'_{ij}(t) = \frac{0.2}{e^{\mu|p(i;t)-p(j;t)|}}. \quad (11)$$

μ is chosen to be equal to 2. The "binding" is done via this second layer. In fact, it is an array of fully connected neurones along with a global controller—global controller defined as in Eqs. (5) and (6). The global controller desynchronises the synchronised neurones from different auditory objects by emitting inhibitory activities whenever there is activity (spikings) in the network [63]. Thus, the network can adapt quickly to input changes (see Table 2 for parameter values).

The selection strategy at the output of the second layer is based on temporal correlation:

- Neurones belonging to the same source synchronise (same spiking phase).
- Neurones belonging to other sources desynchronise (different spiking phase).

8. Masking

Based on the phase synchronisation described in the previous section, a mask is generated by associating zeros and ones to different channels. Energy is normalised in order to have the same sound pressure level (SPL) for all frames. Note that two-source mixtures are considered throughout this article but the technique can be potentially used for more than two sources. In this case, for each time frame n, the labelling of individual channels is equivalent to the use of multiple masks (one for each source).

9. Experiments

9.1. Database and comparison

Martin Cooke's database [10] is used for evaluation purposes. The following intruding signals have been tested: 1 kHz tone, FM siren, white noise trill telephone noise and speech. The aforementioned noises have been added to the target utterance. The results (i.e., audio files) can be found at [43]. Each mixture is applied to our proposed system and the mixed sound sources are separated. The log spectral distortion (LSD) is used as

Table 3
The log spectral distortion (LSD) for four different methods

Intrusion	*SNR* of the initial mixture (dB)	P–R (LSD)	W–B (LSD)	H–W (LSD)	B–R (LSD)
Tone	−2	5.38	18.87	13.59	9.77
Siren	−5	8.93	20.64	13.40	18.94
Tel. ring	3	16.43	18.35	14.05	16.18
White noise	−5	16.82	35.25	26.51	14.84
Male (da)	0	14.92	N/A	N/A	17.70
Female (di)	0	19.70	N/A	N/A	24.04

P–R (our proposed approach), W–B (the method proposed in [63]), H–W (the method proposed in [23]), and B–R (the method proposed in [4]). The intrusion noises are as follows: (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, (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.

Table 4
LSD for two different methods

Mixture	Separated sources	P–R (LSD)	J–L (LSD)	B–R (LSD)
Music and female (AF)	Music	8.01	21.25	14.00
	Voice	17.54	16.49	18.16

P–R (our proposed approach), J–L [27], and B–R [3,4]. The mixture comprises a female voice with musical rock background.

a performance criterion [61,62]. It is defined as[9]

$$LSD = \frac{1}{L}\sum_{l=0}^{L-1}\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left[20\log_{10}\frac{|I(k,l)|+\varepsilon}{|O(k,l)|+\varepsilon}\right]^2}, \quad (12)$$

where $I(k,l)$ and $O(k,l)$ are the FFT of $I(t)$ (ideal source signal) and $O(t)$ (separated source), respectively. L is the number of frames, K is the number of frequency bins and ε is meant to prevent extreme values (equal to 0.001 in our case). We compare our performance with four different approaches proposed in the literature:

- The system proposed in [63] is a neural-network-based CASA system that uses a different neural architecture and a different type of preprocessing.
- The expert-system-based CASA approach proposed in [23], which uses more conventional cues, such as pitch tracking for grouping different sources.
- An approach based on statistical learning from [27]. Note that in the latter case a training is necessary prior to separation.
- A speech enhancement technique based on wavelets by Bahoura and Rouat [3,4].

9.2. Separation performance

Table 3 gives comparison results based on the LSD for a first set of sound files:

- In all cases, our proposed architecture performs better than [63]. Our system causes less distortion of the target signal at the expense of less noise rejection. This is an advantageous strategy in hearing aid design.
- Our proposed technique outperforms [23] when the intrusion is a tone, siren or white noise. For telephone ring, [23] has better scores.
- Our technique outperforms [4] in all cases except for white noise.[10] However, the technique in [4] is a speech enhancement technique that has never been designed for source separation. Speech enhancement techniques are more adapted to background noise (including white noise) removal than to sound source separation tasks. Note that the price to pay for a better performance with our proposed technique, is a higher computational complexity when compared with more conventional speech enhancement techniques such as the one presented in [4].
- For the double-vowel, the LSD has the highest value—showing that separation is more difficult when the interference is speech.

Table 4 shows comparison results based on the LSD for a second set of sound files:

- Sound files used in [27][11] are used for comparison purposes. The current work is compared with two other techniques from [27,4]. B–R gives only one output (processed sound), in contrast with P–R and J–L, which give two extracted sounds for each method (Tables 3 and 4). Therefore, the result of B–R is compared (and the LSD extracted) once with the original music and

[9]For other criteria, such as the PESQ (perceptual evaluation of speech quality) see [54,45].

[10]Even if the LSD is lower for the telephone ring (16.18 instead of 16.43), the perceptive quality for B–R is not as good as that of P–R.

[11]http://home.bawi.org/jangbal/research/demos/rbss1/sepres.html

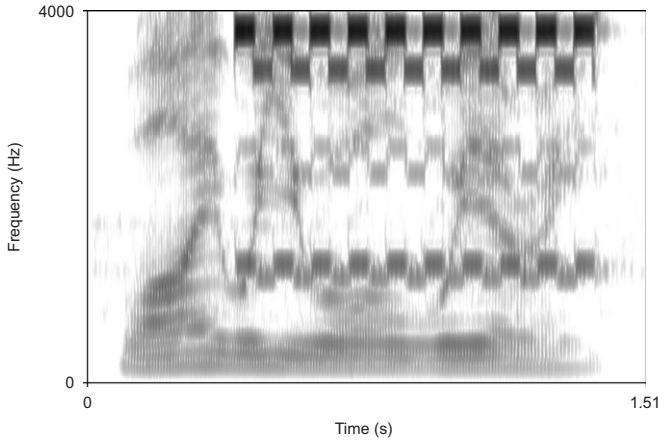

Fig. 4. Mixture of the utterance "Why were you all weary?" with a trill telephone noise.

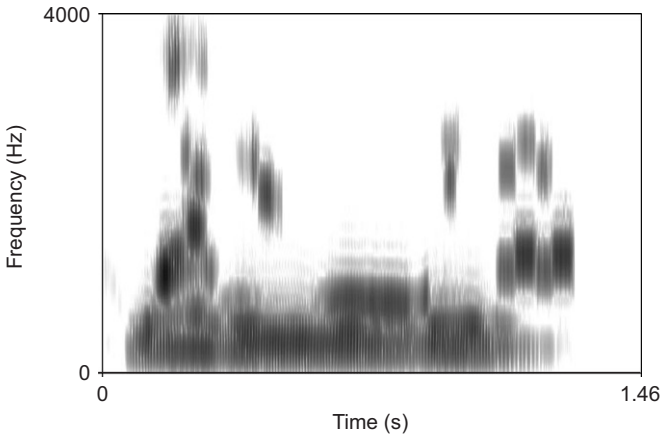

Fig. 6. The synthesised "Why were you all weary?" by the approach proposed in [63] for the trill telephone and utterance mixture.

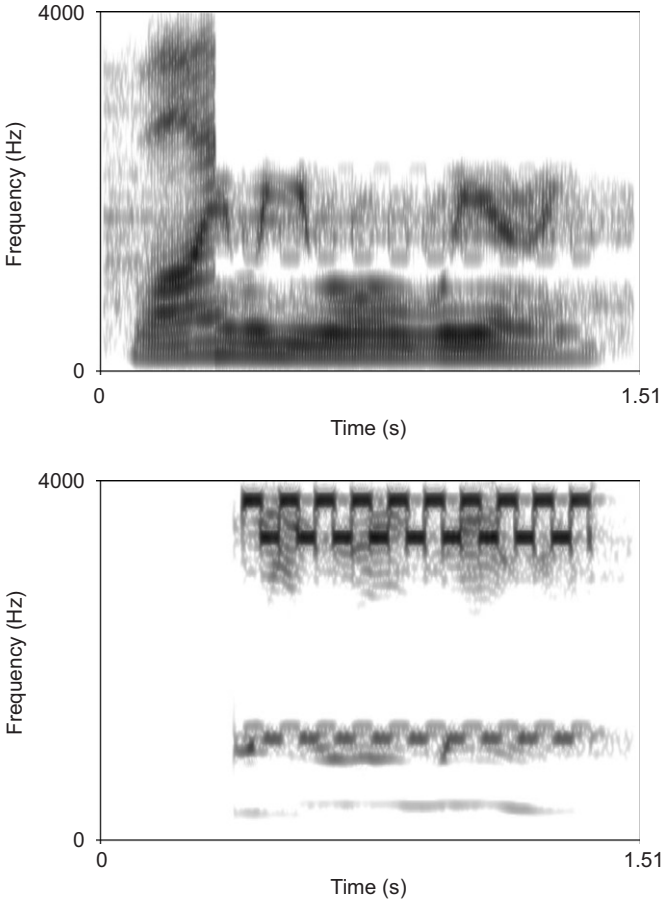

Fig. 5. Top: The synthesised "Why were you all weary?" after the separation by the approach proposed in this article. Bottom: The synthesised trill telephone after the separation by the approach proposed in this article.

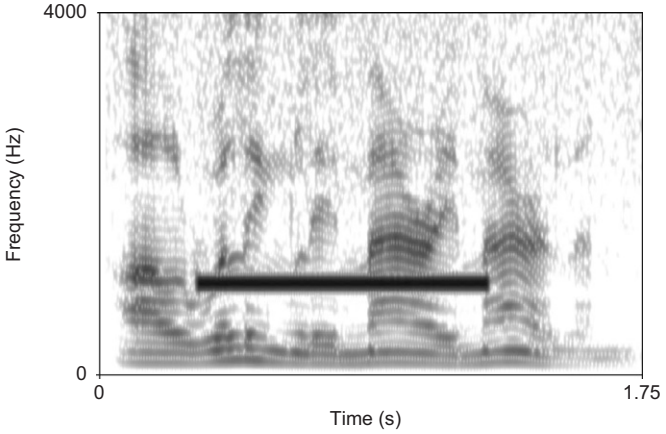

Fig. 7. Mixture of the utterance "I'll willingly marry Marilyn" with 1 kHz tone.

once with the original female speaker. Results from the three techniques are comparable for the extraction of the female speaker voice, while our technique outperforms the other two for the extraction of music from background. Note that our technique does not require any prior statistical training, in contrast with [27].

In the following subsections, spectrograms for different sounds and different approaches are given for visual comparison purposes.

9.3. Separation examples

9.3.1. Separation of speech from telephone trill

Fig. 4 shows the mixture of the utterance "Why were you all weary?" with the telephone trill noise (from Martin Cooke's database). The trill telephone noise (ring) is wideband, interrupted, and structured. Fig. 5 shows the spectrograms of separated utterance and trill telephone, obtained by using our approach. It is interesting to note that the medium- to low-frequency range of the telephone trill has been preserved. Fig. 6 shows the extracted utterance by using [63]. As can be seen, our approach performs better in medium to higher frequencies.

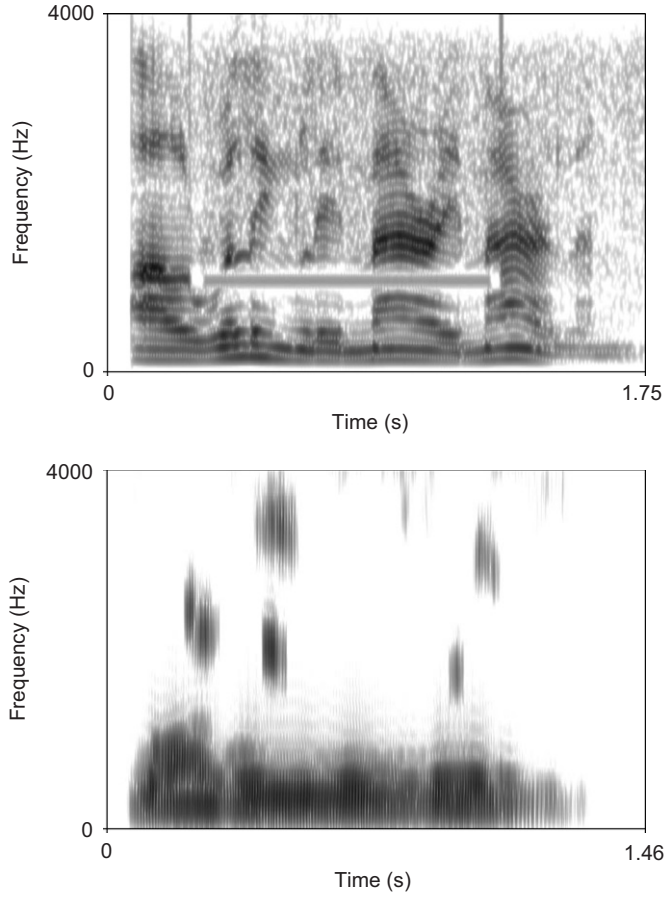

Fig. 8. Top: The separation result for the 1 kHz plus utterance mixture using the approach described in this article. The dynamic range between the darkest gray level and the brightest level is 50 dB. Bottom: The synthesised "Why were you all weary?" by the approach proposed in [63]. The high-frequency information is missing.

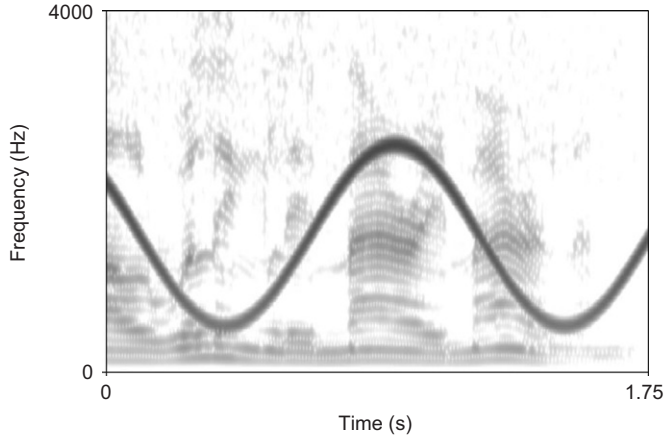

Fig. 9. Mixture of a siren and the sentence "I'll willingly marry Marilyn".

9.3.2. Separation of speech from 1 kHz tone

In this experiment the utterance "I'll willingly marry Marilyn" with a 1 kHz pure tone is used. The tone is narrowband, continuous and structured. Fig. 7 shows the original utterance plus the 1 kHz tone. Fig. 8 shows the

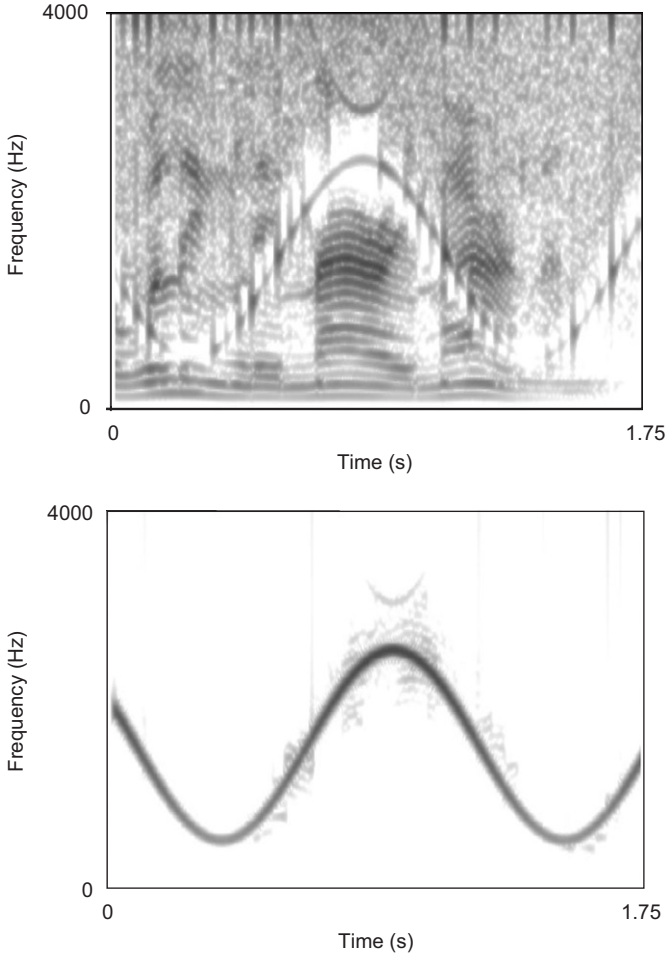

Fig. 10. Top: The separation result for the utterance extraction from the siren plus sentence mixture with our proposed technique. Bottom: The separation result for the siren extraction from the siren plus sentence mixture with our proposed technique.

separation results for our approach and the approach proposed in [63]. The method proposed in [63] removes speech in middle and high frequencies, while these frequencies remain unaffected by our approach. When listening to the signal and according to the LSD (equals to 7.07), the tone has been removed (even if a gray bar is shown on the top panel of Fig. 8.

9.3.3. Separation of speech from an FM signal (siren)

Fig. 9 shows the mixture of the utterance "I'll willingly marry Marylin" with a siren. The siren is a locally narrowband, continuous, structured signal. The bottom panel of Fig. 10 shows the separated siren obtained by our proposed technique. The top panel of Fig. 10 shows the spectrogram of the separated utterance. Fig. 11 shows the spectrogram for the separated utterance using the method proposed in [63].

9.4. Discussions

Other signal-to-noise-ratio (SNR)-like criteria such as the SNR, segmental SNR, PEL (percentage of energy loss),

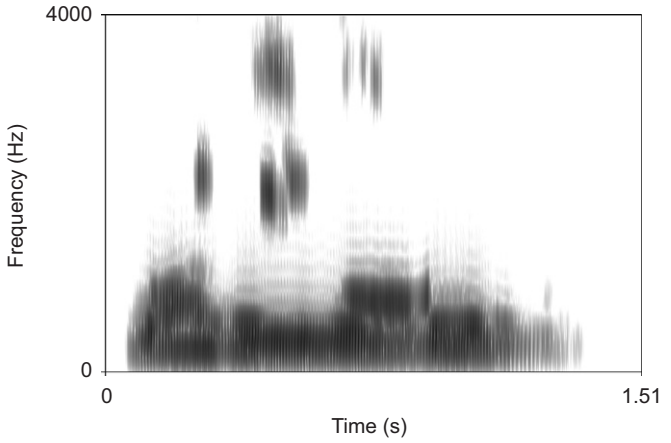

Fig. 11. The synthesised "Why were you all weary?" by the approach proposed in [63] for the siren plus utterance mixture case.

and PNR (percentage of noise residue) are used in the literature in [63,30,22–24,50,44] and can be used as performance scores.

Although criteria like the PEL/PNR, the SNR, the segmental SNR, and the LSD are used in the literature as performance criteria, they do not always reflect the real perceptive performance of a given sound separation system. For example, the SNR, the PEL and the PNR ignore high-frequency information. The LSD does not take into account temporal aspects such as phase. Therefore, LSD will not detect phase distortions in separation results. It is known that performance evaluation cannot be based only on measures such as LSD.[12] Other criteria like the PESQ (perceptual evaluation of speech quality) [54,45] can be used as well for comparison purposes. More investigation should be made to modify the criteria commonly used in source separation.

At any instant of time, the intruding noise such as tones, telephone rings and sirens have narrow bandwidths with strong localised spectral energy. These noises appears easily on the CSM representation. On the other hand, intruding signals such as music and speech are wideband and would not separate from other speech sources based on the CSM, while they do with CAM. CAM and CSM are complementary representations, at least for the limited database used here. Referring back to first sections of the paper, we can write that the separation of two interfering unvoiced speech sounds will very likely not be feasible based only on the CSM and CAM representations. We know from physiology that other representations are available to the auditory system and should be implemented for a more robust system. These new representations should not assume the stationarity of signals, as it has been assumed here by using the Fourier transform.

From the siren interfering signal we observe that the neural network is able to follow the dynamic changes in time and frequency, which is a crucial property of the system. However, the limiting factor is the number of cochlear channels and the width of the sliding window. Audio files with the results for three-source sound separation can be found on the web pages at [43,52].

10. Conclusion

A new system that comprises a perceptive analysis to extract multiple and simultaneous features to be processed by an unsupervised neural network has been proposed. There is no need to tune-up the neural network when changing the nature of the signal. Furthermore, there is no training or recognition phase.

The proposed system has been tested on a limited corpus. Many improvements should be made before considering an extensive use of the approach in real situations. Among them, there is a need for creating new feature maps for onset, offset, etc. Nevertheless, the experiments have led us to the conclusion that computational neuroscience in combination with speech processing offers a strong potential for unsupervised and dynamic speech processing systems.

Even with crude approximation such as binary masking and non-overlapping and independent windows,[13] we obtain relatively good synthesis intelligibility.

Future developments will include the implementation of an efferent feedback control for signal representation adaptation and feature selections as a top-down (schema-driven) processing.

Acknowledgements

The authors would like to thank Philippe Gournay, Hossein Najaf-Zadeh, and Richard Boudreau for reading and commenting the paper, DeLiang Wang and Guoning Hu for their audio files and for fruitful discussions on oscillatory neurones, Christian Feldbauer and Gernot Kubin for their FIR implementation of the gammatone filterbanks and for fruitful discussions on filterbanks and Jean-Marc Valin for discussions on the LSD performance criterion. Many thanks are also due to the anonymous reviewers for their constructive feedbacks.

References

[1] R. Andersen, L. Snyder, D. Bradley, J. Xing, Multimodal representation of space in the posterior parietal cortex and its use in planning movements, Ann. Rev. Neurosci. 20 (1997) 303.
[2] L. Atlas, S.A. Shamma, Joint acoustic and modulation frequency, EURASIP J. Appl. Signal Process. 7 (2003) 668–675.
[3] M. Bahoura, J. Rouat, Wavelet speech enhancement based on the teager energy operator, IEEE Signal Process. Lett. 8 (2001) 10–12.
[4] M. Bahoura, J. Rouat, New approach for wavelet speech enhancement, Eurospeech 2001, Danemark, 2001, pp. 1937–1940.

[12] Sound and demo files are available for listening at: http://www-edu.gel.usherbrooke.ca/picr1601/Demos.htm

[13] Binary masks create artifacts by placing zeros in the spectrum where the interfering source was. The absence of energy at these locations might be heard (hearing of absent signal or musical noise).

[5] F. Berthommier, G. Meyer, Improving of amplitude modulation maps for f0-dependent segregation of harmonic sounds, in: Eurospeech'97, 1997.
[6] S.M. Bohte, H.L. Poutré, J.N. Kok, Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks, IEEE Trans. Neural Networks 13 (2) (2002) 426–435.
[7] A. Bregman, Auditory Scene Analysis, MIT Press, Cambridge, MA, 1990.
[8] G. Brown, M. Cooke, Computational auditory scene analaysis, Comput. Speech Language (1994) 297–336.
[9] M. Cooke, Modelling auditory processing and organisation, Ph.D. Thesis, University of Sheffield (published in the Distinguished Dissertations in Computer Science Series, University of Cambridge Press, paper back, 2005).
[10] M. Cooke, ⟨http://www.dcs.shef.ac.uk/~martin/⟩, 2004.
[11] M. Cooke, D. Ellis, The auditory organization of speech and other sources in listeners and computational models, Speech Commun. (2001) 141–177.
[12] M. Cooke, P. Green, L. Josifovski, A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data, Speech Commun. 34 (2001) 267–285.
[13] D. Ellis, Prediction-driven computational auditory scene analysis, Ph.D. Thesis, MIT, 1996.
[14] F. Gabbiani, H. Krapp, C. Koch, G. Laurent, Multiplicative computation in a visual neuron sensitive to looming, Nature 420 (2002) 320–324.
[15] F. Gaillard, Analyse de scènes auditives computationnelle (CASA): Un nouvel outil de marquage du plan temps-fréquence par détection d'harmonicité exploitant une statistique de passage par zéro, Ph.D. Thesis, INPG, 1999.
[16] W. Gerstner, Spiking neuron models: single neurons, populations, plasticity, Cambridge University Press, Cambridge, 2002.
[17] C. Giguere, P.C. Woodland, A computational model of the auditory periphery for speech and hearing research, J. Am. Statist. Assoc. (1994) 331–349.
[18] S. Grossberg, K. Govindarajan, L. Wyse, M. Cohen, M. ART-STREAM: a neural network model of auditory scene analysis and source segregation, Neural Networks, 2003.
[19] S. Harding, G. Meyer, Multi-resolution auditory scene analysis: robust speech recognition using pattern-matching from a noisy signal, In: EUROSPEECH, September 2003, pp. 2109–2112.
[20] C.K. Henkel, The auditory system, in: D.E. Haines (Ed.), Fundamental Neuroscience, Churchill Livingstone, 1997.
[21] J. Hopfield, Pattern recognition computation using action potential timing for stimulus representation, Nature 376 (1995) 33–36.
[22] G. Hu, D. Wang, Monaural speech segregation based on pitch tracking and amplitude modulation, Technical Report, Ohio State University, 2002.
[23] G. Hu, D. Wang, Separation of stop consonants, in: ICASSP Hong Kong, 2003.
[24] G. Hu, D. Wang, Monaural speech segregation based on pitch tracking and amplitude modulation, IEEE Trans. Neural Networks 15 (2004) 1135–1150.
[25] T. Irino, R. Patterson, Speech segregation using event synchronous auditory vocoder, in: ICASSP, 2003, vol. V, pp. 525–528.
[26] T. Irino, M. Unoki, A time-varying, analysis/synthesis auditory filterbank using the gammachirp, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 98, May 1998, Seattle, Washington, vol. 6, pp. 3653–3656.
[27] G. Jang, T. Lee, Single-channel signal separation using time-domain basis functions, IEEE-SPL (2003) 168–171.
[28] G. Jang, T. Lee, A maximum likelihood approach to single channel source separation, J. Mach. Learn. Res. 4 (2003) 1365–1392.
[29] S. Kim, D.R. Frisina, R.D. Frisina, Effects of age on contralateral suppression of distorsion product otoacoustic emissions in human listeners with normal hearing, Audiol. Neuro Otol. 7 (2002) 348–357.
[30] A.J.W. Kouwe, D.L. Wang, G.J. Brown, A comparison of auditory and blind separation techniques for speech segregation, IEEE Trans. Speech Audio Process. 9 (2001) 189–195.
[31] G. Kubin, W.B. Kleijn, On speech coding in a perceptual domain, in: ICASSP, March 1999, Phoenix, Arizona, vol. 1, pp. 205–208.
[32] N. Levy, D. Horn, I. Meilijson, E. Ruppin, Distributed synchrony in a cell assembly of piking neurons, Neural Networks 14 (6–7) (2001) 815–824.
[33] M. Liberman, S. Puria, J.J. Guinan, The ipsilaterally evoked olivocochlearreflex causes rapid adaptation of the 2f1–f2 distortion product otoacoustic emission, J. Acoust. Soc. Am. 99 (1996) 2572–3584.
[34] C. von der Malsburg, 1981. The correlation theory of brain function, Technical Report Internal Report 81-2, Max-Planck Institute for Biophysical Chemistry.
[35] C. von der Malsburg, The what and why of binding: the modeler's perspective, Neuron (1999) 95–104.
[36] C. von der Malsburg, W. Schneider, A neural cocktail-party processor, Biol. Cybern. (1986) 29–40.
[37] D.K. Mellinger, B.M. Mont-Reynaud, Scene analysis, in: H. Hawkins, T. McMullen, A. Popper, R. Fay (Eds.), Auditory Computation, Springer, New York, 1996, pp. 271–331.
[38] G. Meyer, D. Yang, W. Ainsworth, in: Applying a model of concurrent vowel segregation to real speech, Computational models of auditory function, S. Greenberg, M. Slaney (Eds.), 2001, pp. 297–310.
[39] P. Milner, A model for visual shape recognition, Psychol. Rev. 81 (1974) 521–535.
[40] J. Nix, M. Kleinschmidt, V. Hohmann, Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction, in: EUROSPEECH, September 2003, pp. 1441–1444.
[41] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, M. Allerhand, Complex sounds and auditory images, in: Y. Cazals, L. Demany, K. Horner (Eds.), Auditory Physiology and Perception, Pergamon Press, Oxford, 1992, pp. 429–446.
[42] J. Pena, M. Konishi, Auditory spatial receptive fields created by multiplication, Science 292 (2001) 249–252.
[43] R. Pichevar, ⟨http://www-edu.gel.usherbrooke.ca/picr1601/⟩ 2007.
[44] R. Pichevar, J. 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.
[45] R. Pichevar, J. Rouat, A quantitative evaluation of a bio-inspired sound segregation technique for two- and three-source mixtures sounds, in: Lecture Notes in Computer Science, Springer, Berlin, 2004, vol. 3445, pp. 430–435.
[46] R. Pichevar, J. Rouat, C. Feldbauer, G. Kubin, A bio-inspired sound source separation technique in combination with an enhanced FIR gammatone analysis/synthesis filterbank, in: EUSIPCO Vienna, 2004.
[47] F. Plante, G. Meyer, W. Ainsworth, Improvement of speech spectrogram accuracy by the method of reassignment, IEEE Trans. Speech Audio Process. (1998) 282–287.
[48] M.J. Reyes-Gomez, B. Raj, D. Ellis, Multi-channel source separation by factorial HMMs, in: ICASSP, 2003.
[49] M. Riesenhuber, T. Poggio, Are cortical models really bound by the binding problem?, Neuron 84 (1999) 87–93.
[50] N. Roman, D. Wang, G. Brown, Speech segregation based on sound localization, J. Acoust. Soc. Am. (2003).
[51] D.F. Rosenthal, H.G. Okuno (Eds.), 1998. Computational Auditory Scene Analysis, L. Erlbaum.
[52] J. Rouat, ⟨http://www.gel.usherbrooke.ca/rouat/⟩ 2005.
[53] J. Rouat, Y.C. Liu, D. Morissette, A pitch determination and voiced/unvoiced decision algorithm for noisy speech, Speech Commun. 21 (1997) 191–207.
[54] J. Rouat, R. Pichevar, Source separation with one ear: proposition for an anthropomorphic approach, EURASIP J. Appl. Signal Process. (9) (2005) 1365–1374.

[55] S. Roweis, Factorial models and refiltering for speech separation and denoising, in: Eurospeech 2003.
[56] S.T. Roweis, One microphone source seperation, in: NIPS, Denver, USA, 2000.
[57] H. Sameti, H. Sheikhzadeh, L. Deng, R. Brennan, HMM based strategies for enhancement of speech signals embedded in nonstationary noise, IEEE Trans. Speech Audio Process. (1998) 445–455.
[58] J.L. Schwartz, P. Escudier, Auditory processing in a post-cochlear neural network: vowel spectrum processing based on spike synchrony, in: EUROSPEECH, 1989, pp. 247–253.
[59] P. Tang, J. Rouat, Modeling neurons in the anteroventral cochlear nucleus for amplitude modulation (AM) processing: application to speech sound, in: Proceedings of the International Conference on Spoken Language Processing, October 1996, p. Th.P.2S2.2.
[60] N. Todd, An auditory cortical theory of auditory stream segregation, Network Comput. Neural Syst. 7 (1996) 349–356.
[61] J.-M. Valin, F. Michaud, J. Rouat, D. Ltourneau, Robust sound source localization using a microphone array on a mobile robot, in: IEEE/RSJ-International Conference on Intelligent Robots and Systems, October 2003.
[62] J.-M. Valin, J. Rouat, F. Michaud, Microphone array post-filter for separation of simultaneous non-stationary sources, in: IEEE International Conference on Acoustics Speech Signal Processing, 2004.
[63] D. Wang, G.J. Brown, Separation of speech from interfering sounds based on oscillatory correlation, IEEE Trans. Neural Networks 10 (3) (1999) 684–697.
[64] D. Wang, D. Terman, Image segmentation based on oscillatory correlation, Neural Comput. 9 (1997) 805–836.
[65] M. Weintraub, A theory and computational model of auditory monaural sound separation, Ph.D. Thesis, Stanford, 1985.

Ramin Pichevar was born in March 1974, in Paris, France. He received his bachelor of science degree in electrical engineering (electronics) in 1996 and his master of science in electrical engineering (telecommunication systems) in 1999, both in Tehran, Iran. He received his Ph.D. in electrical and computer engineering from Université de Sherbrooke, Québec, Canada in 2004. During his Ph.D., he taught courses on signal processing and computer hardware as a lecturer. In 2001 and 2002, he performed two summer internships at Ohio State University (USA) and at the University of Grenoble (France), respectively. From November 2004 to July 2006, he was a postdoctoral and research associate in the computational neuroscience and signal processing laboratory at the University of Sherbrooke under an NSERC (Natural Sciences and Engineering Council of Canada) Idea to Innovation (I2I) grant. He is presently a research scientist at the Communications Research Centre (CRC), Ottawa, Canada and an adjunct professor at the University of Sherbrooke. His domains of interest are signal processing, Computational Auditory Scene Analysis (CASA), neural networks with emphasis on bio-inspired neurons, speech recognition, audio and speech coding, sparse coding, digital communications, discrete-event simulation, and image processing.

Jean Rouat holds a master degree in Physics from Université de Bretagne, France (1981), an E.&E. master degree in speech coding and speech recognition from Université de Sherbrooke (1984) and an E.&E. Ph.D. in cognitive and statistical speech recognition jointly with Université de Sherbrooke and McGill University (1988). From 1988 to 2001, he was with 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 from UQAC. From September 2006 to March 2007, he was with McMaster University with the ECE department. He is now with Université de Sherbrooke where he founded the Computational Neuroscience and Intelligent Signal Processing Research group. Since February 2007, he is also invited professor in the biological sciences dept from Université de Montréal. His research interests cover audition, speech and signal processing in relation with networks of spiking neurons. 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, etc.). He is a senior member of the IEEE and was on the IEEE Technical Committee on Machine Learning for Signal Processing from 2001 to 2005.