Acoustic Representation and Processing: It is time!

Jean ROUAT *
NECOTIS, GEGI
Universit de Sherbrooke

Stéphane LOISELLE
NECOTIS, GEGI
Universit de Sherbrooke

Ramin PICHEVAR
Advanced Audio Systems
CRC, Ottawa

Abstract

From physiology we learn that the auditory system extracts simultaneous features from the underlying signal, giving birth to simultaneous representations of audible signals. The use of the Rank Order Coding has also been recently hypothesized in the mammalian auditory system. In a first application we compare a very simplistic speech recognition prototype that uses the Rank Order Coding with a conventional Hidden Markov Model speech recognizer. In a second application, we combine a simultaneous auditory images representation with a network of oscillatory spiking neurones to segregate and bind auditory objects for acoustical source separation. We discuss on the importance of the time in acoustic processing.

1 Introduction

This paper has two essential goals: *i)* emphasizes on the importance of time in the features extracted by the auditory system – with illustration of the influence of such time structures on performance of a simple prototype speech recognizer and on a sound source separation system – and *ii)* brings up the question: "How far can we go in sound source separation systems when not taking into account prior knowledge?"

In the present paper we present two speech processing systems based on an auditory motivated approach. One system is a prototype speech recognizer that uses the sequence of spikes, as a recognition feature, in a network of spiking neurons. When compared with a Hidden Markov Models recognizer, preliminary results show that the approach as a great potential for transients recognition and situations where no sufficient data are available to train the speech recognizer. The second system shows that separation and segregation of simultaneous audio sources can be performed by using the temporal correlation paradigm to group auditory objects coming from the same source. This system does not need prior knowledge of the sources.

1.1 Suitable representations

We know that, for speech processing, an efficient and suitable signal representation has first to be found. Ideally the representation should be adapted to the problem in hand and has not to be the same for both systems (for exmple, the separation and recognition cues are different). With at least two interfering speakers and voiced speech, it is observed that the separation is relatively easy – speakers with different fundamental frequencies – since spectral representations or auditory images exhibit different regions with structures dominated by different pitch trajectories. Hence, amplitude modulation of cochlear filter outputs (or modulation spectrograms) are discriminative. In situations where speakers have similar pitches, the separation is difficult and features, like the phase, should be preserved by the analysis to increase the discrimination. Most conventional source segregation or recognition systems use an analysis that is effective, as long as speech segments under the analysis

*Emails: Jean.Rouat@usherbrooke.ca, Stephane.Loiselle@usherbrooke.ca, Ramin.Pichevar@usherbrooke.ca. J. Rouat is visiting McMaster Univ., R. Pichevar is now with the Communications Research Centre, Ottawa.

window are relatively stationary and stable. Furthermore, these systems need a sufficient amount of data to be trained. For segregation with at least 2 simultaneous speakers and overlapping unvoiced speech segments, a signal-dependent and fine analysis is required. In this situation, separation and recognition cannot be performed with conventional speech analysis that are usually based on spectral and time-averaged features. It is very likely that this kind of system would fail. Hence, adaptive (or at least multi-features) and dynamic signal analysis is required.

1.2 So, which signal representation?

The question is still open, but the signal representation should enhance speech structured features that ease discrimination and differentiation.
Based on the literature on auditory perception, it is possible to argue that the auditory system generates a multi–dimensional spatio-temporal representation of the one dimensional acoustic signal. In the first half of the paper (speech recognition application) we use a simple time sequence of 2D auditory images (cochlear channels – neuron thresholds) or the time sequence of 3D Shamma's multiscale analysis (tonotopic frequency – temporal rate – modulation of the auditory spectrum). In the second half of the paper we use 2D auditory image (AM modulation and spectrum of cochlear envelops) sequences and we treat them as videos.

2 Perceptive Approach

From physiology we learn that the auditory system extracts simultaneous features from the underlying signal, giving birth to simultaneous time structured multi-representations of speech. We also learn that fast and slow efferences can selectively enhance speech representations in relation to the auditory environment. This is in opposition with most conventional speech processing systems that use a systematic analysis [1] that is effective only when speech segments under the analysis window are relatively stationary and stable.

2.1 Physiology: Multiple Features

Inner and outer hair cells establish synapses with efferent and afferent fibres. The efferent projections to the inner hair cells synapse on the afferent connection, suggesting a modulation of the afferent information by the efferent system. On the contrary, other efferent fibres project directly to the outer hair cells, suggesting a direct control of the outer hair cells by the efferences. It has also been observed that all afferent fibres (inner and outer hair cells) project directly into the cochlear nucleus. It has a layered structure that preserves frequency tonotopic organisation where one finds very different neurons that response to various features [2]. In the inferior colliculus of the cat, Schreiner and Langner [1, 2] have shown that there exists a highly systematic topographic representation of AM parameters. Maps showing best modulation frequency have been determined. The pioneering work by Robles, Ruggero and Evans [3] [4][5] reveals the importance of AM-FM [3] coding in the peripheral auditory system along with the role of the efferent system in relation with adaptive tuning of the cochlea. Small neural circuits in relation with *wideband inhibitory input* neurons are observed by Arnott *et al.* [6] in the cochlear nucleus. These circuits, explain the response of specialised neurons to frequency position of sharp spectral notches.
It is also known that the auditory efferent system plays a crucial role in enhancing signals in background noise [7] [8] [9]. Kim *et al.* [9] measure the effect of aging on the medial olivocochlear system and suggest that the functional decline of the medial olivocochlear system with age precedes outer hair cell degeneration.
It is clear from physiology that multiple and simultaneous representations of the same input signal are observed in the cochlear nucleus [10] [11]. In the remaining parts of the paper, we call these representations, *simultaneous auditory images*.

---

[1] A systematic analysis extracts the same features independently of the signal context. Frame by frame extraction of Mel Frequency Cepstrum Coefficients (MFCC) is an example of a systematic analysis.

[2] onset, chopper, primary–like, etc.

[3] Other features like transients, ON, OFF responses are observed, but are not implemented here.

2.2 Rank Order Coding and Temporal Correlation

VanRullen et *al.* [12] discuss the importance of the relative timing in neuronal responses and have shown that the coding of the Rank Order of spikes can explain the fast responses that are observed in the human somatosensory system. Along the same line, Nätschlager and Maass [13] have technically shown that information about the result of the computation is already present in the current neural network state long before the complete spatio-temporal input patterns have been received by the neural network. This suggests that neural networks use the temporal order of the first spikes yielding ultra-rapid computation in accordance with the physiology by [14, 12]. As an example, we can cite the work by DeWeese et *al.* [15], where the authors observe in the auditory cortex of the rat that transient responses in auditory cortex can be described as a binary process, rather than as a highly variable Poisson process. Once again, these results suggest that the spike timing is crucial. As the Rank Order Coding is one of the potential neural codes that respects these spike timing constraints, we explore here a possible way of integration in the context of speech recognition.
On the other hand, we explore, in the context of source separation, the use of the *Temporal Correlation* to compute dynamical spatio–temporal correlation between features obtained with a bank of cochlear filters. This time the coding is made through the synchronization of neurons. Neurons that fire simultaneously will characterize the same sound source.

3 Exploration in Speech Recognition

3.1 Rank Order Coding

Rank Order Coding has been proposed by Simon Thorpe and his team from CERCO, Toulouse to explain the impressive performance of our visual system [16, 17]. The information is distributed through a large population of neurons and is represented by spikes relative timing in a single wave of action potentials. The quantity of information that can be transmitted by this type of code increases with the number of neurons in the population. For a relatively large number of neurons, the code transmission power can satisfy the needs of any visual task [16]. There are advantages in using the relative order and not the exact spike latency: the strategy is easier to implement, the system is less subject to changes in intensity of the stimulus and the information is available as soon as the first spike is generated.

3.2 System Overview

In this work, we use the temporal order of spikes at the output of two auditory models. One is a very simple peripheral auditory channels representation (which we call *cochlear channels* model) while the second is a far more complete model that includes mid–brain and cortical representations of sounds (which we call *complete* model).
We also use two models of neurons. In the first model, spikes are obtained with simple fixed thresholds (which we call *threshold* model), without integration. Different neurons can have a different internal threshold value. The second model is the conventional leaky integrate and fire representation of neurons (LIAF) with different internal thresholds.

3.3 Speech Analysis Modules

Auditory models The first is a model of peripheral auditory system which is crudely modelled by a gammatone filter-bank followed by rectification and compression with a square-root law. The second model has been proposed by Shamma and his team [18] and his a complete model from the periphery to the auditory cortex. Spike trains are generated by feeding the neuronal models (described in next subsections) with one of the two analysis modules described here.

Neuronal models *Simple threshold neuronal model:* For each signal envelope, from each cochlear channel, we use three neurons with different thresholds (denoted 1, 2 and 3). A spike is generated when the signal envelope exceeds one of the internal neuron's threshold. After producing a spike, a neuron becomes inactive for the remaining time of the stimulus. With such model we only capture the first instant for which the envelope signal is sufficiently high and we ignore the other spikes. *The Leaky Integrate and Fire neuronal model:* Our simple Integrate-and-Fire neuron model has four parameters: adaptive threshold, leaky current, resting potential and reset potential. Throughout the

simulation, the neuron integrates the input to determine the neuron's internal potential evolution. Once the internal potential is sufficient (it reaches the threshold), a spike is generated.

3.4 Illustration with the simple analysis and thresholding

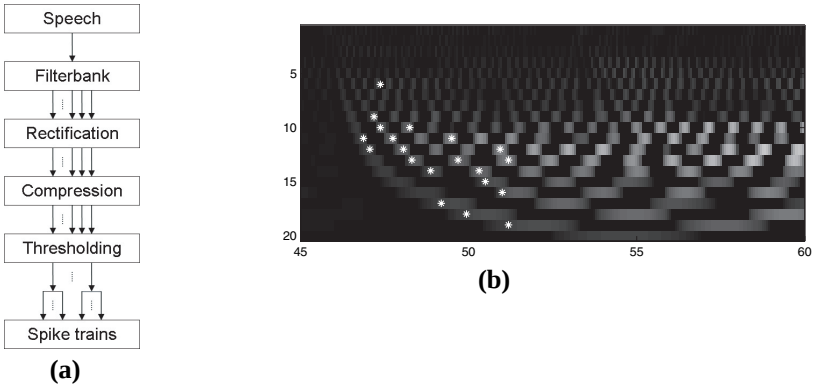


Figure 1: Illustration of spike sequence generation on a French digit 'un' [ɛ̃] with the cochlear channel analysis and threshold neurons. **(a)** The signal is first filtered through a cochlear gammatone filter-bank. It is then rectified and compressed with a square-root law. For each channel, three threshold neurons are used with different threshold values. If the amplitude in the channel reaches one of the neuron's threshold, a spike is generated. After firing, a neuron becomes inactive for the remaining time of the stimulus. **(b)** White stars represent spikes. The x-axis is the time samples (sampling frequency of 16 kHz) and the y-axis shows the filter-bank channels. Center frequencies of channels 1 and 20 are respectively equal to 8000 and 100 Hz.
.

We illustrate the sequence generation and processing with the cochlear channels analysis used in conjunction with the threshold neuronal models. Figure 1, page 4, sumarizes the preprocessing and the generation of spikes with the threshold neuronal model.

3.5 Learning and Recognition Modules

The same training and recognition procedure has been used with the simple thresholding and leaky integrate and fire neuronal models. During training a template is generated for each reference word. It is a sequence of the first most likely N cochlear channel/threshold numbers.

3.6 Experiments

Speech Database We performed a proof-of-concept test of speech recognition using in house speech databases made of 10 French digits spoken by 5 men and 4 women and of the French vowels spoken by the same 5 men and 5 women (including the 4 women from the digits database). Each speaker pronounced ten times the same digits (from 0 to 9) and vowels. The speakers were presented with random sequences of digits and vowels to be read.

Training and Recognition For each digit (or vowel), two reference models are used for the recognizer (one pronunciation for each sex). Recognition has been performed on all pronunciations of each speaker. During recognition and for a given digit (or vowel), only two speakers were represented in the reference models. Experiments have been conducted with two combinations: cochlear filter–bank analysis with simple threshold neurons (noted as Cochlear–Threshold) and the complete auditory model with LIAF neurons (noted as Complete–LIAF).

Reference System A conventional MFCC and Hidden Markov Model speech recognizer has been trained with the same training set than with the ROC prototype [4].

[4]The same reference pronunciation has been used for each digit (or vowel).

3.7 Recognition scores with limited training data

Table 1: Averaged recognition rates on the five French vowels [aəioy] for the HMM speech recognizer, the Cochlear–Treshold and Complete-LIAF speech recognizers

HMM	Cochlear–Treshold	Complete-LIAF
90%	89%	94%

Recognition of vowels The 5 French vowels [aəioy] recognition has been made with three recognizers. The conventional reference HMM, one ROC prototype with the cochlear analysis and the threshold neuron model and another ROC prototype with the complete auditory model with LIAF neurons. The average recognition rates are reported on table 1. The HMM and the Cochlear–Treshold recognizers have comparable performance while the complete model (Complete-LIAF) is better. The Cochlear–Treshold recognizer does relatively well (recall that it uses only one spike per channel/threshold neuron) even if it uses only the first signal frames. On the opposite, the HMM and the Complete-LIAF systems use the full vowel signal.

French digit recognition We report in table 2 the results for the very simple system (Cochlear–Treshold) and those from the MFCC–HMM recognizer.

Table 2: Recognition for each pronunciation of the ten French digits – Cochlear filter analysis combined with the one time threshold neuron (Cochlear–Threshold system) and MFCC with HMM.

Cochlear–Threshold	Total Recognition Rate: 65 %									
Digits	1	2	3	4	5	6	7	8	9	0
Scores (%)	93	76	64	75	46	75	13	75	60	67

MFCC–HMM	Total Recognition Rate: 52 %									
Digits	1	2	3	4	5	6	7	8	9	0
Scores (%)	16	61	46	36	90	80	33	5	49	100

3.8 Discussion

Both ROC based speech recognizer prototypes do surprisingly well when compared with a state of the art HMM recognizer.
The Complete–LIAF system uses the same length of the test signal than the HMM system but without the stationary assumption of the MFCC analysis and yields better results on the vowels when the training set is limited (only 1 reference speaker for each sex). The simple Cochlear–Threshold system has comparable results than the HMM on stationary signals (vowels) and much better results on the consonants of the digits. Clearly, one reference speaker per sex with only one occurrence is not sufficient to train the HMM. Furthermore, the transient cues from the consonants are spread out with the stationary MFCC analysis.
Apart from the fact that the Rank Order Coding scheme could be a viable approach to the recognition it is important to notice that with only one spike per neuronal model (reported results with the Cochlear–Threshold system) the recognition is still promising. It is interesting to link these preliminary results with the arguments of Thorpe and colleagues [14] [19]. They argue that first spike latencies provide a fast and efficient code of sensory stimuli and that natural vision scene reconstructions can be obtained with very short durations. If such a coding is occuring in the auditory system, the study conducted here could be a good start point in the design of a speech recognizer.
The statistical HMM speech recognizer and our ROC prototypes are complementary. The ROC prototypes seem to be robust to transient and unvoiced consonants recognition (which is known to be difficult for statistical based speech recognizers) while the HMM based systems are know to be very robust for stationary voiced speech segments. Also, an important aspect to consider is that the HMM technology dates from the middle of the seventies and plenty of good training algorithms are available which is not yet the case for the ROC. For now, a mixed speech recognizer, that relies *i)* on a Perceptive & ROC approach for the transients and *ii)* on the MFCC & HMM approach for the voiced segments could be viable. While the HMM recognizer performance can be improved by increasing the training set, the performance of our prototypes could benefit from various pre- or post-processing schemes that are currently under investigation.

4 A sound source separation system based on synchronisation of neurons

In this part of the paper we present a work that explores the feasibility of sound source separation without any prior knowledge on the interfering sources. In other words we begin to pave the road to answer to the question: "How far can we go in the separation of sound sources without integrating any priori knowledge on the sources?".
We assume here that sound segregation is a generalised classification problem, in which we want to bind features – extracted from the auditory image representations – in different regions of a neural network map.

4.1 Proposed System Strategy

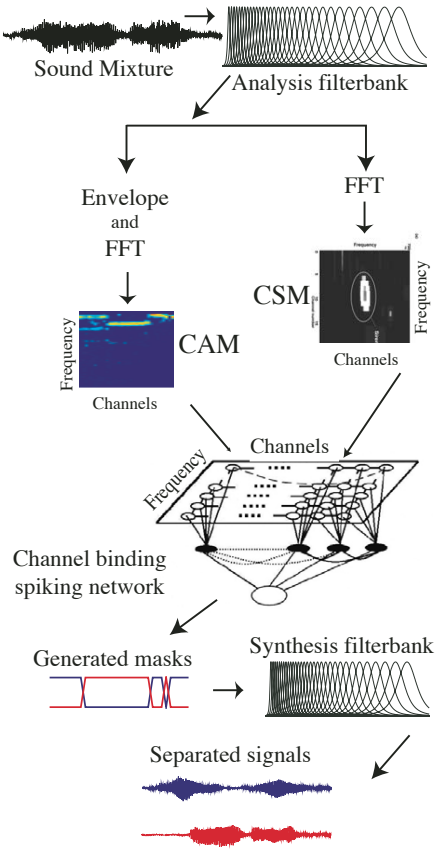
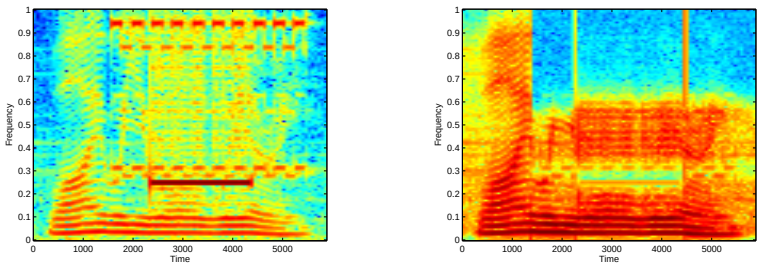


Figure 2: Source Separation System. Two auditory images are simultaneously generated – a Cochleotopic/AMtopic map (CAM) and a Cochleotopic/Spectrotopic map (CSM) – in two different paths. Based on the neural synchrony of the lower layer neurons, a binary mask is generated to mute – in time and across channels – the synthesis filterbank's channels that do not belong to the desired source. The first layer of the neural layer segregates the auditory objects on the CAM/CSM maps and the second layer binds the channels that are dominated by the same stream. For now, the user decides which representation (CAM or CSM) is being presented to the neural network.

The sound mixture is processed by an enhanced FIR Gammatone filterbank that mimics the behavior of the cochlea [20]. From the output of the cochlear channels two different anthropomorphic maps are generated (figure 2). The amplitude modulation map, that we call Cochleotopic/AMtopic (CAM) Map and the Cochleotopic/Spectrotopic Map (CSM) that encodes the averaged spectral energies of the cochlear filterbank output. These maps partially mimic the behavior of the peripheral auditory pathway. These maps are based on the computation of the FFT (Fast Fourier Transform) and envelope detection [21] (figure 2).
The first representation somehow reproduces the AM processing performed by multipolar cells

(Chopper-S) from the anteroventral cochlear nucleus [11], while the second representation is closer to the spherical bushy cell processing from the ventral cochlear nucleus [10].
For now, 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. 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 could be used to suppress the interference (or separate all sources with adaptive masks).
Our separation technique uses the Computational Auditory Scene Analysis that is based on the computational implementation of ideas exposed by Bregman [22]. A two-layered network of spiking neurons is used to perform cochlear channel selection (figure 2) based on temporal correlation: neurons associated to those channels belonging to the same sound source synchronize. A more detailed description of the system is given in [23, 24]. It has been tested on two-sources and three-sources sound source separation situations. Results can be found in [23, 24, 25]. Different criteria such as PEL (Percentage of Energy Loss), PNR (Percentage of Noise Reduction), LSD (Log Spectral Distortion), and PESQ (Perceptual Evaluation of Speech Quality) have been used for the evaluation of the quality of separation. According to these criteria, it has been shown [23, 24, 25] that the proposed system outperforms other state of the art systems that do need to be trained to know the source characteristics.



(a) **(b)**

Figure 3: Illustration of the separation on three simultaneous sources. (a) The spectrogram of the mixture of an utterance, a tone, and a telephone ring. (b) The spectrogram of the extracted utterance. The telephone and the tone have been suppressed, while the voice has been preserved. Results can be heard and evaluated on one of the authors' web page: [26] [27].

4.2 Discussion

The uses a perceptive analysis to extract multiple and simultaneous time structured features to be processed by an unsupervised neural network. 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. Even with a crude approximation such as binary masking, non overlapping and independent time window, we obtain relatively good synthesis intelligibility.

5 Conclusion

Starting in the mid '80s, auditory models preserving time organized features were already proposed and tested on corrupted signals but with limited success because of pattern recognizers' inabilities to exploit the rich time-structured information generated by these models. Spiking neural networks open doors to new systems with a stronger integration between analysis and recognition. The understanding of the performance of the ROC (with little training data) in relation with Bayesian learning methods is certainly an interesting research issue for the future. It is also important to note that acceptable monophonic source separations can be performed without prior knowledge by using the temporal correlation. Of course, when possible, a more robust system would combine both Bayesian learning and temporal correlation.

Aknowledgments

Many thanks to Simon Thorpe and Daniel Pressnitzer for receiving S. Loiselle during his 2003 summer session in CERCO, Toulouse. The authors would also like to thank DeLiang Wang and Guoning Hu for fruitful discussions on oscillatory neurons, Christian Feldbauer, Gernot Kubin for discussions on filterbanks and software exchanges. This work has been funded by NSERC and Université de Sherbrooke. S. Loiselle has been funded by FQRNT of Québec for the year 2006.

References

[1] Cristophe E. Schreiner and John V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. I.. the anterior auditory filed (AAF). *Hearing research*, 21:227–241, 1986.

[2] C.E. Schreiner and G. Langner. Periodicity coding in the inferior colliculus of the cat. II, topographical organization. *Journal of Neurophysiology*, 60:1823–1840, 1988.

[3] Luis Robles, Mario A. Ruggero, and Nola C. Rich. Two-tone distortion in the basilar membrane of the cochlea. *Nature*, 349:413, Jan. 1991.

[4] E. F. Evans. Auditory processing of complex sounds : An overview. In *Phil. Trans. Royal Society of London*, pages 1–12, Oxford, 1992. Oxford Press.

[5] Mario A. Ruggero, Luis Robles, Nola C. Rich, and Alberto Recio. Basilar membrane responses to two-tone and broadband stimuli. In *Phil. Trans. Royal Society of London*, pages 13–21, Oxford Press, 1992.

[6] Robert H. Arnott, Mark N. Wallace, Trevor M. Shackleton, and Alan R. Palmer. Onset neurones in the anteroventral cochlear nucleus project to the dorsal cochlear nucleus. *JARO*, 5(2):153–170, 2004.

[7] C. Giguere and Philip C. Woodland. A computational model of the auditory periphery for speech and hearing research. *JASA*, pages 331–349, 1994.

[8] M.C. Liberman, S. Puria, and J.J. Jr. Guinan. The ipsilaterally evoked olivocochlearreflex causes rapid adaptation of the 2f1-f2 distortion product otoacoustic emission. *JASA*, 99:2572–3584, 1996.

[9] S.H. 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. *Audio. Neuro Oto.*, 7:348–357, 2002.

[10] C. K. Henkel. The Auditory System. In Duane E. Haines, editor, *Fondamental Neuroscience*. Churchill Livingstone, 1997.

[11] P. Tang and J. Rouat. Modeling neurons in the anteroventral cochlear nucleus for amplitude modulation (AM) processing: Application to speech sound. In *Proc. Int. Conf. on Spok. Lang. Proc.*, Oct 1996.

[12] Rufin VanRullen, Rudy Guyonneau, and Simon J. Thorpe. Spike times make sense. *Trends in Neurosciences*, 28(1):4, January 2005.

[13] Thomas Natschläger and Wolfgang Maass. Information dynamics and emergent computation in recurrent circuits of spiking. In *NIPS*, December 2003.

[14] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[15] Michael DeWeese, Michael Wehr, and Anthony Zador. Binary spiking in auditory cortex. *The Journal of Neuroscience*, 23(21):7940–7949, 2003.

[16] Rufin VanRullen and Simon J. Thorpe. Surfing a spike wave down the ventral stream. *Vision Research*, 42(23):2593–2615, August 2002.

[17] Laurent Perrinet. *Comment déchifrer le code impulsionnel de la Vision ? Étude du flux parallèle, asynchrone et épars dans le traitement visuel ultra-rapide.* PhD thesis, Université Paul Sabatier, 2003.

[18] Shibab Shamma. Physiological foundations of temporal integration in the perception of speech. *Journal of Phonetics*, 31:495–501, 2003.

[19] S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14(6-7):715–725, 2001.

[20] R. Pichevar, J. Rouat, C. Feldbauer, and G. Kubin. A bio-inspired sound source separation technique in combination with an enhanced FIR gammatone Analysis/Synthesis filterbank. In *EUSIPCO*, 2004.

[21] R. Pichevar and J. Rouat. Cochleotopic/AMtopic (CAM) and Cochleotopic/Spectrotopic (CSM) map based sound source separation using relaxation oscillatory neurons. In *IEEE NNSP, Toulouse*, 2003.

[22] Al Bregman. *Auditory Scene Analysis*. MIT Press, 1994.

[23] J. Rouat and R. Pichevar. Source separation with one ear: Proposition for an anthropomorphic approach. *EURASIP Journal on Applied Signal Processing*, 1365–1373, June 2005.

[24] R. Pichevar, J. Rouat, C. Feldbauer, and G. Kubin. A bio-inspired sound source separation technique in combination with an enhanced FIR gammatone Analysis/Synthesis filterbank. In *EUSIPCO*, 2004.

[25] Ramin Pichevar and Jean Rouat. A Quantitative Evaluation of a Bio-Inspired Sound Source Segregation Technique Based for two and three Source Mixtures. In G. Chollet, A. Esposito, M. Faundez-Zanuy, and M. Marinaro, editors, *Advances in Nonlinear Speech Modeling and Applications*, volume 3445 of *Lectures Notes in Computer Science*, pages 430–434. Springer Verlag, 2005.

[26] Ramin Pichevar. http://www-edu.gel.usherbrooke.ca/pichevar/Demos.htm, 2004.

[27] J. Rouat. http://www.gel.usherb.ca/rouat, 2004.