Acoustic Representation and Processing: It is time!

JEAN ROUAT†‡, STÉPHANE LOISELLE†, RAMIN PICHÉVAR†*
† Université de Sherbrooke, Canada. ‡ visiting McMaster univ., Canada. * CRC, Ottawa, Canada

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 neurons to segregate and bind auditory objects for acoustical source separation. We discuss on the importance of the time in acoustic processing.

The auditory system: Multiple simultaneous features
Physiology: Multiple and simultaneous representations of the same input signal are observed in the cochlear nucleus [1]. We call these representations, simultaneous auditory images.

Rank Order Coding (ROC) and Temporal Correlation
The Rank Order of spikes can explain the fast responses that are observed in the human somatosensory system [2]. Natschläger and Maass [3] have shown that the information about the order of the computation is already present in the current neural network state before the complete spatio-temporal input patterns have been received by the neural network. Some neural networks use the temporal order of the first spikes yielding ultra-rapid computation in accordance with the physiology by [4, 2].

Example of sequence and weight generation
Example of spikes generation with a simple cochlear filterbank [6]

Source Separation System [8]

Source Separation System with two auditory images simultaneously generated. The first layer of the network segregates the auditory objects and the second layer binds the channels that are dominated by the same stream. Sound demonstration at [9].

Discussion
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 (after separation) [9].

Conclusion
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.

References

Acknowledgments
A. Thorpe, D. Pressnitzer for receiving S. Loiselle in CERCO, D. Pressnitzer and S. Loiselle for discussions on audiovisual and software exchanges. NSERC and Université de Sherbrooke.

Accepted at Quebec for the year 2006.

Example of similarity computation with a neuron already tuned to the channel sequence (13, 12, 14, 15, 11, 10, 9, ...), inhibition increasing in time.

Speech Recognition
Database, learning & reference system
– French digits spoken by 5 men and 4 women and French vowels spoken by the same 5 men and 5 women. Each speaker pronounced ten times the same digits and vowels. For each digit (or vowel), 2 reference models are used for the recognizer.
– Recognition performed on all pronunciations of each speaker.
– A Hidden Markov Model based system has been trained on the same training set (seventeen cepstral coefficients for each time frame) [6].

Vowels Recognition
Averaged recognition rates on the five French vowels (a[øy] for the HMM, the Cochlear–Threshold and Complete-LIAF recognizes

Digits Recognition
Ten French digits; Cochlear filter analysis combined with the one time threshold neuron. Complete-LIAF systems now a first signal frame. HMM and Complete-LIAF systems use full signal.

Recognition for ten French digits; MFCC with HMM

Table Rank Order Coding

Dendrite
Example of similarity computation with a neuron already tuned to the channel sequence (13, 12, 14, 15, 11, 10, 9,...), inhibition increasing in time.

Source Separation: Temporal Correlation
– A Hidden Markov Model based system has been trained with the Cochlear–Threshold system) the recognition results with the Cochlear–Threshold system) the recognition
– The simple Cochlear–Threshold system has better results for the digits of the digits. The transient cues from the consonants are spread out with the stationary MFCC analysis.
– With only one spike per neuronal model reported results with the Cochlear–Threshold system the recognition is promising.
– It is interesting to link this with the arguments of Thorpe and colleagues [4] [7]: First spike latencies provide a fast and efficient code of sensory stimuli.

Conclusion
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.

References

Acknowledgments
A. Thorpe, D. Pressnitzer for receiving S. Loiselle in CERCO, D. Pressnitzer and S. Loiselle for discussions on audiovisual and software exchanges. NSERC and Université de Sherbrooke.

Accepted at Quebec for the year 2006.

Example of sequence and weight generation
Example of sequence and weight generation

Sequence generation on a French digit ‘un’ [1] with a cochlear channel analysis and threshold neurons. (a) 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.

Example of sequence and weight generation
Example of sequence and weight generation

Sequence with the spike order of the first 20 cochlear channel threshold numbers to produce a spike and generated weights k:

Source Separation: Temporal Correlation
How far can we go in sound source separation with no prior knowledge on the sources? We explore the use of the Temporal Correlation to compute dynamical spatio-temporal correlations between features from auditory images. The coding is made through the synchronization of neurons. Neurons that fire simultaneously characterize the same sound source.

Source Separation: Temporal Correlation
How far can we go in sound source separation with no prior knowledge on the sources? We explore the use of the Temporal Correlation to compute dynamical spatio-temporal correlations between features from auditory images. The coding is made through the synchronization of neurons. Neurons that fire simultaneously characterize the same sound source.