

# Spatio-temporal Pattern Recognition with Neural Networks: Application to speech

Jean ROUAT<sup>1,2</sup>

<sup>1</sup> ERMETIS, Sciences Appliquées, Université du Québec à Chicoutimi\*\*\*

<sup>2</sup> Neuro-Heuristique, I.P., Faculté de Médecine, Université de Lausanne

**Abstract.** The processing or the recognition of non stationary process with neural networks is a challenging and yet unsolved issue. The paper discuss the general pattern recognition framework using neural networks in relation with the understanding of the peripheral auditory system. We propose a short-time structure representation of speech for speech analysis and recognition. We give examples of neural networks architecture and applications that are designed to take into account the time structure of the process to be analysed.

## 1 Introduction

Most of the contemporary pattern processing or recognition techniques assume that the pattern or the time series to be recognised are stationary. Furthermore, in real life applications, the information is most of the time corrupted, partial or noisy (image, speech, etc.). Therefore, the pattern recognisers have also to be robust.

## 2 Speech and the Auditory System

Speech is a good example of a complex signal that is very difficult to analyse or recognise by computers when produced in real life situations. The best contemporary technology is not able to propose speech recognition systems with acceptable performance when the speech is corrupted or noisy.

### 2.1 Structures of Speech

One of the reason is that speech is a fuzzy structured signal. It is structured in reference to the production and hearing systems that impose the constraints and structures on the speech (formants, voiced/unvoiced, spectral and temporal distributions, etc.). For speech scientists those constraints and structures seem to be somewhat fuzzy when analysing the signal without any a priori knowledge about the production and hearing systems.

---

\*\*\* Jean\_Rouat@uqac.quebec.ca

## 2.2 Where is the Recognition Performed?

Another reason is that the perceptive system does not process speech as pattern recognition systems usually do. To a certain extent, it is true that the cochlear nucleus, the superior olivary complex and the colliculus, for example, are apparently specialised and they might perform 'signal processing' tasks. But there is evidence that a partial 'recognition' is already made at the level of the intermediate auditory system. In fact, the time constants and the best modulation transfer functions of the cortical auditory neurons reveal that the time locking to periodicities with frequencies greater than 100Hz is very rarely observed in the auditory cortex. This is lower to pitch and formant frequencies. Therefore, pitch and periodic stimulus are probably processed at the peripheral or intermediate level of the auditory system. Periodic stimulus with different frequencies can elicit similar evoked potential responses in the auditory cortex. On the opposite, the response of the cortical neurones is very strong when the stimulus are transients. The afferent pathways to the primary auditory cortex have preserved the neural timing of transient stimulus events. In the auditory cortex, neural activity patterns in response to tones show a low-frequency periodic response with strong ON and OFF responses. Those responses are the result of a collaborative work between hundred of cells. The global activity has a complex dynamic that encode the information.

The auditory system seems to process speech by taking into account the specific speech structure and the recognition task is probably performed at all stages of the auditory processing (continuum) in contrast with standard pattern recognisers where the parameter extraction is first made and then the recognition is performed.

## 2.3 Short-time Structures of Speech

We define the short-time structures of speech as the characteristics of speech observed through a peripheral auditory system on very short-time scales (a few ms). We are interested in the short-time structure observed at the output of a cochlear filter bank in terms of amplitude modulation (AM). This structure is typical of speech. Generally speaking, the structures depend very much on the way speech is produced. The production and hearing systems are closely related and adapted to each others. The hearing system provides time structured representations of speech that are representative and characteristic of the speech production modes. These structures can not be observed via conventional speech analysis techniques and are important to speech analysis and recognition.

## 3 Modulation in the Auditory System: an Example of Structure

The modulation information is one of the main cues extracted by the auditory system. Various work is made regarding the physiology of AM processing

[2] [4] [7] [8]. For example, Schreiner and Langner [4] [8] show that the inferior colliculus of cats contains a highly systematic topographic representation of AM parameters and maps showing 'best modulation frequency' have been determined. Shannon [9] reports experiments on listeners that can understand artificial speech. This artificial speech is obtained by preserving the envelope and deleting the fine structure of the original speech in three adjacent spectral bands.

For the nerve fibres whose characteristic frequency (CF) is close to a formant frequency, a phase-coupling to the formant frequency or to an adjacent harmonic is observed with little or no envelope modulation as the discharge pattern of the fibre is dominated by a single large harmonic component. Other fibres may show modulations corresponding to harmonic interactions.

Today, most of the speech recognisers extract acoustical parameters from the signal based on spectral representations of speech. By doing so, one has to assume that the signal is stationary under the analysis window. This yields a representation of speech that is an estimate of the time-averaged parameter values. Therefore, the short-time structure of speech is partially hidden by the analysis and the AM fine structure can not be seen.

### 3.1 The Short-term AM Structure

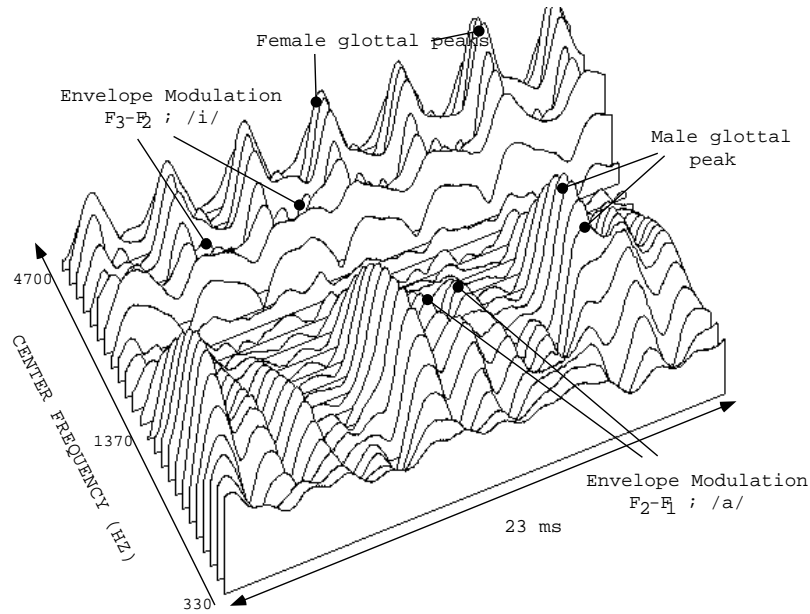
An example of a short-term AM structure is given below. A bank of 24 filters centred on 330Hz to 4700Hz is used. The output of each filter is a bandpass signal with a narrow-band spectrum centred around  $f_i$  where  $f_i$  is the central frequency (CF) of channel  $i$ . The output signal  $s_i(t)$  from channel  $i$  can be considered to be modulated in amplitude and phase with a carrier frequency of  $f_i$ .

$$s_i(t) = A_i(t)\cos[\omega_i t + \phi_i(t)] \quad (1)$$

$A_i(t)$  is the modulating amplitude (envelope),  $\phi_i(t)$  is the modulating phase and  $\omega_i = 2\pi f_i$ .

An image representation of the structure can be obtained by plotting  $A_i(t)$  versus time and central frequency. The x axis is the time and the y axis is expressed in Hertz according to the ERB (equivalent rectangular bandwidth) scale [5]. The image colour is the  $A_i(t)$  variable.

Fig.1. shows the envelope structure for a signal segment comprising 23 ms of speech. A female pronounced the letter /i/ along with a male speaker saying /a/. In that example, a segregation of the speakers is possible when looking to the structure along the time dimension (x axis) and across the channels (y axis). The low frequency channels (centre frequency for channels 1-13: 330-1500 Hz) are dominated by the signal from the male speaker (/a/) while the high frequency channels are dominated by the female voice (/i/). The most interesting modulations occur during the glottal explosion or just after. The figure includes approximately 3 glottal explosions from the male voice and 7 glottal explosions for the woman. Modulations due to harmonics interactions occur during a pitch period just after the glottal explosion.



**Fig. 1.** Envelopes for a two speakers speech segment: /a/ from a male speaker, /i/ from a female speaker.

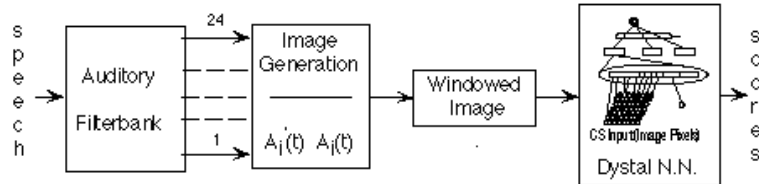
### 3.2 The Recognition

Previous sections introduced a short-term AM structure of speech that can not be observed by using conventional analysis techniques. To perform the recognition, neural networks have to take into account the internal time structure of the representation. We propose two examples on speech. The first uses an associative neural network memory for a speech recognition task and the second is based on an oscillatory neural network for a speaker verification task.

**The Associative Neural Network: Dystal.[1]** Dystal (DYNAMICALLY STABLE Associative Learning) is proposed by Alkon et *al.* and is inspired from a marine snail and from the hippocampus of a rabbit. We assume that clusters are characterised by an explicit encoding of the reference patterns in dendritic patches of neurons [1]. The network associatively learns correlations and anticorrelations between time events occurring in pre synaptic neurons. Those neurons synapse on the same element of a common post synaptic neurone. A learning rule modifies the cellular excitability at dendritic patches. These synaptic patches are postulated to be formed on branches of the dendritic tree of vertebrate neurons. Weights are associated to patches rather than to incoming connection. After learning, each patch characterises a pattern of activity on the input neurons.

In comparison with most commonly used networks, the weights are not used to store the patterns and the comparison between patterns is based on normalised

correlations instead of projections between the network input vectors and the neurone weights. Based on the Dystal network, a prototype vowel recognition system has been designed and preliminary results show that the short-time AM structure carries information that can be used for recognition of voiced speech [6].



**Fig. 2.** Architecture of a speech recogniser prototype.

An image representation is obtained by plotting the product  $A_i(t) \cdot A_i(t)'$  versus time and central frequency. That representation enhances transitions and amplitude modulations of the envelope. A sliding and synchronised to glottal peak window is moved on the image representation of speech and is used as input to the associative memory. A preliminary experiment has been successfully conducted on four vowel clusters: /a/, /i/, /y/ and /ε/. Detailed results are presented in [6].

**The Oscillatory Neural Network: a Novelty Detector.** The topology of the network is inspired from cortical layer IV. The neurone model is based on the integrate and fire model. Each neurone receives excitatory or inhibitory inputs from their neighbourhood. A global inhibitor neurone is connected to each neurone in the network. It regulates and stabilises the network activity. Synaptic rules are used to adapt the weights of the local and global connections.

There is no difference between 'training' and recognition. 'Training' is always performed except when the network is stable. As soon as signals are presented to the network, it learns by modifying the synaptic weights. During training, the network oscillates. When changes on weights are too small to modify the dynamic of the network, we assume that 'training' and recognition have been completed. The network is then declared to be in a stable state. The time necessary to reach that state is the relaxation time. It is used as the novelty detection criteria.

The clusters are not explicitly encoded. The relaxation time of the network is used to characterise the input pattern. A short relaxation time implies that the pattern has been already 'seen' by the network. This paradigm allows the creation of novelty detection systems with a high degree of robustness against noise. Such network is used for the recognition of noisy numbers [3] and experiments on speech are currently made with larger networks. For example, a speaker verification system is being designed. The network uses the short-term AM structure to verify if the speaker belongs to the 'authorised' cluster.

## 4 Conclusion

The recognition of non stationary spatio-temporal process is very difficult when using formal neural networks. A better understanding of the processing of information in the brain should ease the introduction of new paradigms in pattern recognition with Neural Networks.

We gave two examples with applications to speech that are based on the short-term AM structure of speech. One application uses an associative memory for vowel recognition while the other uses a novelty detector for speaker verification. Those applications are recent and are still under development in order to evaluate their viability in relation with real life applications.

## Acknowledgements

This work has been supported by the NSERC of Canada and by the fondation from Université du Québec à Chicoutimi. Many thanks to Yves de Ribaupierre and Alessandro Villa for stimulating discussions.

## References

1. Alkon, D.L., Blackwell, K.T., Barbour, G.S., Rigler, A.K. and Vogl, T.P.: Pattern-recognition by an artificial network derived from biologic neuronal systems. *Biological Cybernetics*, vol. **62** (1990) 363–376
2. Frisina, R. D. et al.: Differential encoding of rapid changes in sound amplitude by second-order auditory neurons. *Exp. Brain Res.*, vol. **60** (1985) 417–422
3. Ho, T. V. and Rouat, J.: A Novelty Detector using a Network of Integrate and Fire Neurons. *ICANN97*.
4. Langner, G. and Schreiner, C.E.: Periodicity coding in the inferior colliculus of the cat. Neuronal mechanisms. *Journal of Neurophysiology*, vol. **60**, 6, (1988) 1799–1822
5. Patterson, R.D.: Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, vol. **59**, 3 (1976) 640–654
6. Rouat, J. and Garcia, M.: A prototype speech recogniser based on associative learning and nonlinear speech analysis. In Proc. of the Workshop on *Computational Auditory Scene Analysis*, International Joint Conference (IEEE-ACM) on Artificial Intelligence, (1995) 7–12. To be published in, *Readings In Computational Auditory Scene Analysis*, Edited by H. Okuno and D. Rosenthal, Erlbaum.
7. Schreiner, C. E. and Urbas, J. V.: Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research*, vol. **21** (1986) 227–241
8. Schreiner, C.E. and Langner, G.: Periodicity coding in the inferior colliculus of the cat. Topographical organization. *Journal of Neurophysiology*, vol. **60**, 6 (1988) 1823–1840
9. Shannon, R. V et al.: Speech Recognition with Primarily Temporal Cues. *Science*, vol. **270** (1995) 303–304