

## PROCESSING OF NOISY PATTERNS WITH A CONNECTIONIST SYSTEM USING A TOPOGRAPHIC REPRESENTATION OF SPEECH

J. Leboeuf, D. Bérroule, LIMSI(CNRS), BP 30, 91406, Orsay-cedex

### ABSTRACT

This paper is aimed at presenting some preliminary results of a word recognition experiment, where the signal to be identified has been affected by added speech. The notion of topographic representation, upon which the underlying connectionist model is based, will be first introduced. A method is then propounded to obtain such a representation for speech, by mapping a continuous spectrum onto a bit pattern.

### INTRODUCTION

A memory model has been developed, in which speech recognition consists in the selective propagation of signals along pathways defined inside a network of processing units (ref 1) (ref 2). For the purpose of processing real speech, we were led to design a method for providing a spectral representation compatible with the model working constraints.

### FROM CODED TO TOPOGRAPHIC REPRESENTATION

From the recognition device landmark, the acoustic signal appears as a continuous waveform, occurring at a given memory location commonly referred to as the system's input. The pattern to be recognized is thus initially entirely defined by its variable and continuous amplitude  $a(t)$ , not by the spatial coordinates  $(x, y, z)$  of its container, which can be set to any constant values  $(X, Y, Z)$ . If we represent a memory unit by square brackets indexed by its coordinates in the memory field, the input signal can be expressed by:  $[a(t)]_{X,Y,Z}$ . This signal accounts for many different sounds propagated in the system's environment, among which some are considered as "noise", because they disturb the identification of one particular speech sound. The most difficult case happens when a speech signal is modified by added speech. The usual way to deal with that problem in practical applications mainly consists in the use of directional and close-talking microphones; in laboratory experiments, noise may also be artificially suppressed thanks to sound-proof rooms.

A theoretical solution would be to transform  $a(t)$  into a larger number of variables, some of which are associated with the signal/under focus, the others being imputable to noise. If a variable is a position inside the recognizer's memory, a variable corresponding to noise might be easily occulted during the recognition process, thanks to higher level informations. However, signal representations that have been proposed till now are amplitude-dependent, in the sense that they involve a limited number of parameters, the variable values of which are all significant; this means that all registers containing these values participate in recognition, independently of their memory location. It might be the cause of difficulties encountered in noisy environments.

An attempt is made to resolve this problem through the transformation of continuous speech, defined anywhere by its continuous and variable amplitude over time, into a signal defined by its discrete and variable location in time and inside a three-dimensional field.

## A METHOD OF SPEECH ENHANCEMENT

The Fourier Transform may be considered as a mean to increase the dimensionality of  $a(t)$ :

$$[ a(t) ]_{X,Y,Z} \implies [ a(\delta t, \delta f) ]_{X,Y,Z} \quad (1)$$

This is the usual *coded representation* of speech.

The following formulation is aimed at expressing the 2-step signal transformation that is propounded. Each elementary signal provided by the Fourier Transform is first assigned to a variable and discrete location  $\delta x$  along a spatial dimension. Speech enhancement then transforms  $a(t)$  into the non-significant  $A(t)$  which can possibly propagate toward a variable position  $\delta x, \delta y, \delta z$  through the network:

$$[ a(t) ]_{X,Y,Z} \implies [ a(\delta t) ]_{\delta x, Y, Z} \implies [ A(\delta t) ]_{\delta x, \delta y, \delta z} \quad (2)$$

With the previous coded representation, a variable was the content of a given memory subset. In that case, the recognition task consisted in comparing templates, in order to determine the best match between a reference code and an unknown code. With the *topographic representation*, a variable is a given signal which can occur in different locations; the task is to find out the value of a given variable by making the unknown signal propagate toward a specific memory subset.

In the case of many superimposed signals, the transformation gives:

$$[ \sum_i a_i(t) ]_{X,Y,Z} \implies \Delta_i( [ A(\delta t) ]_{\delta x_i, \delta y_i, \delta z_i} ) \quad (3)$$

where  $\Delta_i$  stands for the union of  $i$  memory subsets, minus their intersection, where signals may inhibit each others. For the intersection to be minimized, each  $A_i(\delta t)$  must have enough room to be represented, which calls for a high definition in both temporal and spatial dimensions. This is the reason why we first implemented a signal processing structure involving two parallel and complementary analyses (ref 3). However, the word recognition which is reported here has been carried out according to the following most simple schema:

- The spectral analysis is performed by Fourier transform applied to a pre-emphasi signal which is viewed through a hamming window; it delivers N elementary signals assigned to the network's input.
- The corresponding N feature-detectors are distributed along the frequency scale, and connected to each other through lateral links, according to the *mexican hat function*: each unit receives the weighted activation of its closer neighbours and the weighted inhibition of other more distant neighbours. That *lateral inhibition* gives rise to competition between detectors. As a result, the spectrum is condensed over its maxima, a first step toward the desired topographic representation.

- Each feature-detector responds only to spectral variations, and remains inhibited during a subsequent refractory period. This so called *short-term adaptation* is aimed at detecting discrete events over time, such as spectral onsets, and thus participate in spectral enhancement.
- As the resulting elementary signals propagate through the network, their shape becomes less and less significant, while the pathways taken lead to characteristic locations. The loss of amplitude-related information is thus balanced by a gain in position-related information. The decision is based on those output pathways (word detectors) which are most activated.

It may be remarked that using lateral inhibition and short-term adaptation operators (fig. 1) does not obviously constitute the only possible way to increase the topographic and discrete components of a spectral representation. For instance, using a channel-by-channel waveform modeling, a signal can be decomposed into elementary waveforms discretely repartitioned in the time-frequency plane (ref 4).

### SOME PRELIMINARY RESULTS

- Processing of theoretical topographic representations of speech.

Simulated topographic representations of syllables /RI/ and /PA/ were shown to the software simulation (ref 5) of the connectionist system. The first time they were shown, two pathways leading to syllable detectors were created inside the most peripheral layer of the network. The second time, the words /PAPA/ and /PARI/ were trained, thus inducing two other paths inside a deeper layer. It is shown that the system can cope with time distortions of events. Removing a given amount of events (up to 60%) from the learned representation does not induce misrecognition. If the representation of syllable /RI/ is superimposed on the /PA/ that begins the word /PARI/ to be recognized, propagations first occurs in parallel along the two paths corresponding to /RI/ and /PA/, and then along the path associated with /PARI/ inside the deeper layer. Here we had neglected the interaction between representations, leading to the cancelation of part of the representation to be identified. This will not be the case in the following experiment.

- A word recognition experiment

The goal of this experiment was to compare the performances of our system (N = 64 detectors) with those of a DTW algorithm working on classical parameters (8 MFCC). The corpus was made of 10 sequences of the ten digits; for the purpose of noisy pattern recognition, the signal of the number "ten" has been added to each element of 2 digit sequences; the signal beginnings have been automatically synchronized. Their S/N ratio has been adjusted to 0 db. Results given in (4) show the best performances of the classical device  $S_1$  in a simple recognition task, while our system  $S_2$  seems to be more efficient when submitted to noisy patterns.

	noise free	noisy
$S_1$	100/100	6/20
$S_2$	98/100	18/20

(4)

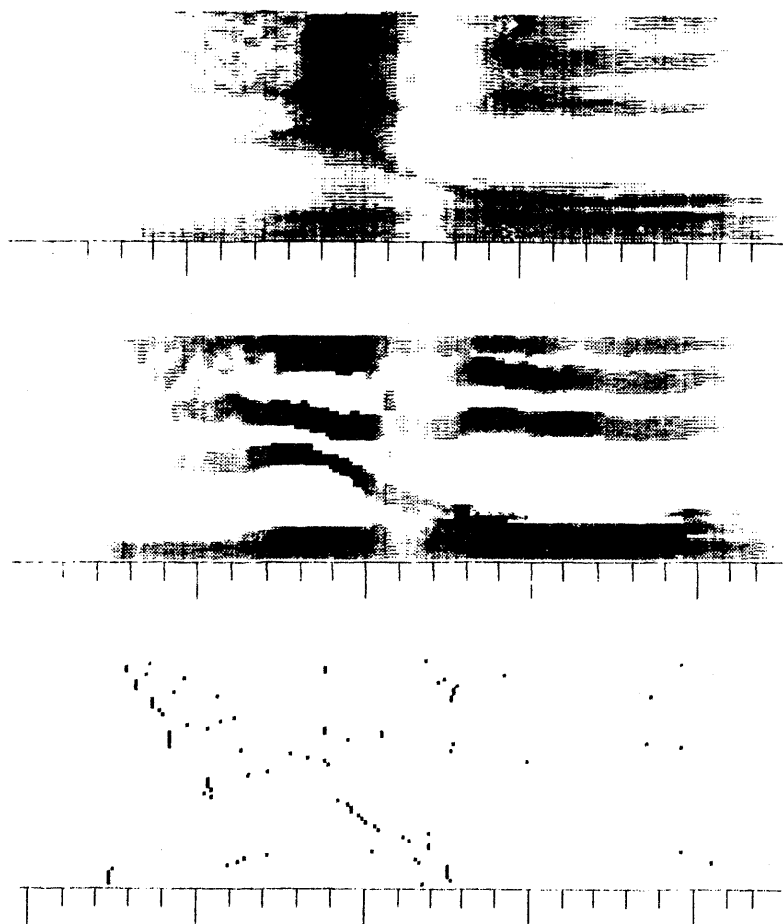


Figure 1. The signal corresponding to digit "zero" is analysed with a FFT (spectrum at the top), passed through a lateral inhibition operator (middle), and through a short-term adaptation operator (bottom).

#### REFERENCES

1. D Béroule, Un modèle de mémoire adaptative, dynamique et associative pour le traitement automatique de la parole ( Thèse 3ème cycle, Orsay, mai 1985)
2. J Leboeuf, Un Système Connexionniste pour le Traitement Automatique de la Parole ( Thèse 3ème cycle, Orsay, to appear in 1987)
3. D Béroule, The Adaptative, Dynamic and Associative Memory model: a possible future tool for vocal computer-human communication ( The Structure of Multimodal Dialogue, M.M.Taylor, F.Neel, D.G.Bouwhuis (Eds.) Amsterdam, North-Holland, 1987)
4. J S Lienard, Speech Analysis and Reconstruction using short-time, elementary waveforms (ICASSP IEEE, Dallas, 1987)
5. J Leboeuf, Présentation d'une simulation du modèle ADAM ( LIMSI's report 86-2, Orsay, 1986)