MANAGEMENT OF TIME DISTORTIONS THROUGH ROUGH COINCIDENCE DETECTION

Dominique G. Béroule

LIMSI/CNRS BP.133, 91403 ORSAY cedex

ABSTRACT

This paper presents a parallel processing principle implemented in a network of processing units, which requires a specific spectral representation of speech. This so-called topographic representation can be obtained through non-linear transformations that are described. Robustness to time variations of the resulting speech recognition system is shown to result from the duration of the internal signals that are synchronized and summed by the processing units.

1. INTRODUCTION

When trying to apply an information processing system to real-world items, appropriate transformations have to be performed. Current pre-processing methods, which code speech into a sequence of vectors, are well adapted to the sequential treatment of numbers by computers. But another type of machine, based on different processing principles, would not necessarily require similar transformations.

For instance, the parallel processing principle presented in this paper cannot work on the output of a classical speech analyzer. Instead of a sequence of continuous parameters, speech is more appropriately considered here as a parallel flow of discrete spectral events.

2. A PRINCIPLE OF PARALLEL PROCESSING

Guided Propagation is a parallel processing principle inspired by psychophysiological data, and implemented inside an associative memory (1). The initial aim of this research work was to investigate a system in which any given information would be represented by an active position in memory. This so-called topographic representation, as opposed to the coded representation of the computer's memory, must be associated with a data retrieval process.

In terms of activation, this is the process through which the memory position corresponding to the current memory input can be activated. Applied to the identification of patterns, this process must also be robust to input variations. Moreover, if the position to be assigned to a given information is not provided, an unsupervised learning method is required.

The proposed solution to these problems can be viewed as searching for the part of memory that is activated simultaneously by the different flows of information involved. For this purpose, a set of pathways which convey the activation issued from the input (stimuli), is intersected with a set of internal pathways which support the spontaneous activation of memory (expectations). Every flow of activation feeds in parallel a subset of memory, and it is the intersection of the different subsets which satisfies both the expectations of the system and the current combination of stimuli. This point of intersection can easily be retrieved, since it is the most activated one. In order to detect it, threshold units are brought in. Set in every point of intersection between the bundles of pathways, they are regulated so as to respond when their input signals coincide.

Accordingly, the way data are stored in this memory consists in forcing both spatial and temporal coincidence of signals (2). Spatial coincidence is obtained by directing a group of spectral stimuli and the contextual signal to the same processing unit. The corresponding connections are created in the course of processing, taking context into account. Temporal coincidence (synchronization) is obtained by associating a time-delay with each input to a unit.

To sum up, Guided Propagation (Fig. 1) can be implemented in a network of processing units (Fig. 2), the main role of which being to detect coincidence between the system's expectations and actual stimuli.

3. PATTERN CLASSIFICATION THROUGH GUIDED PROPAGATION

In classical approaches, Pattern Classification is usually expressed in terms of the partition of feature space into mutually exclusive regions. An input pattern is thus assigned a single vector in a N-dimensional feature space; the region surrounding a given vector is the class to which its associated pattern belongs. Accordingly, a speech sample is usually represented by a single N-dimensional vector, not only in classical recognition systems, but also in connectionist systems such as the Multi-Layer Perceptron (2)(3). A reference vector is associated with each region of the feature space during a preliminary training process. Then, recognition is based on a measure of similarity between each of these references and the input pattern.

Similarity is estimated using a certain global distance between vectors.

With the approach we propose, each pattern class is represented by a memory pathway. Each reference pathway responds to a spatio-temporal configuration of stimuli, taking into account their possible variations in position. An unknown pattern feeds in parallel the reference pathways, thus contributing to their spontaneous activation. Stimuli are collected by processing units distributed along the pathways. Similarity is estimated by these units, in the form of a coincidence rate. If the input matches the spatio-temporal configuration associated with the i-th path better than the others, then the input is classified as from the i-th pattern class.
Here, the input consists of many 2-dimensional (time and frequency) vectors.

A class of patterns is defined as the union of the elementary decision areas associated with each vector (see Fig. 3). These areas correspond to the elementary receptive fields of the processing units which respond to this class of patterns.

The proposed representation exhibits the following theoretical properties:
- whatever its intensity, variability occurs in no more than two dimensions, and is divided between many vectors; along the time axis, only their relative positions are significant;
- different clusters can share the same elements; a single extra vector may differentiate two of them;
- the similarity criteria requires a subset of the vectors, large enough for reaching a threshold. It is therefore not necessary that every elementary area contain a vector;
- since they are independent, every vector can belong to the representation of a given signal among superimposed ones.

This property has been checked in an experiment on noisy digit recognition, which showed better results with Guided Propagation than with a classical algorithm (5).

4. SIGNAL PROCESSING FOR THE DETECTION OF DISCRETE EVENTS

In order to guide propagation in memory, speech events must satisfy the following constraints:

1/ Events must be distributed over time and at least a spatial dimension of memory. Using the Fourier Transform, the speech signal can be decomposed into many elementary components. Signals issued from this analysis can be mapped onto a set of memory inputs called Feature Detectors. In this way, each Feature Detector is assigned a frequency channel within a spectral representation.

| parameters: size of the time-window |

2/ At any given instant, the pathway preferably taken by the internal flow depends on the localized activity of a few frequency channels. When many of them are activated, it is also the case for many pathways; the broader the instantaneous activation, the more ambiguous the input pattern and the less selective the guiding of propagation. In order to minimize the number of activated channels, competition is implemented between them. Feature Detectors are connected to each other through lateral links, according to the Mexican Hat Function (6). Thus, each unit receives the weighted activation of its closest neighbours and the weighted inhibition of other more distant neighbours. The global effect of this so called Lateral Inhibition is a spectrum enhancement, resulting in a condensing over its maxima. The location of these maxima is improved, as an offset to the loss of fine amplitude variations, and in agreement with the topographic representation.

Parameters: range of inhibitory and excitatory connections, weighting coefficients.

3/ Within a given frequency channel, the events that are detected must also be localized and be robust to time distortions. Now, remarking that the strongest time variations concern the steady sounds of speech such as vowels and fricatives, the detection of events inside them is avoided. Events are only detected in transient parts of the spectrum, in the form of spectral onsets and offsets. Whatever the duration of a steady sound is, the channel is kept inhibited during a certain refractory period, after the onset that caused the detection of an event until a comparable offset. This operator, close to physiological Short-Term Adaptation (7), performs a kind of time compression inside each independent channel. It can be easily implemented with a Variability Function and a detection threshold.
The first occurrence of a speech item causes the creation of a pathway in memory. Any new occurrence may exhibit the following differences with its reference:

- **mixing events.** The detection of events being based on competition between frequency channels together with temporal inhibition, some of the known events may be inhibited. The similarity decision depends on the proportion of the total number of events.

- **extra events.** They do not hold up the traffic of the internal flow, and thus do not hinder the recognition process. They are taken into account by the learning process.

- **distorted events.** Variations of the events position may happen in both frequency and time dimensions.

In classical approaches, temporal variations of speech are usually expressed by variations of the parameters that represent it. By comparing the parameter values, the $N^{th}$ vector of a given pattern can be found to match better with the $(N+i)^{th}$ or $(N-j)^{th}$ vector of the reference pattern. In the representation proposed here, time variations appear in a more straightforward way, as variations of spectral events along the time axis.

In Dynamic Time Warping, spectral parameters are organized into a sequence of vectors; only the global variations of instantaneous spectra are considered: the distance between reference and unknown vectors is a mixed compound of the local distances between each parameter. This kind of global variability can also be found in our representation, in the form of shifts of event subsets along the time dimension. Local variations also occur within each group of events.

With regard to group variations, long-lasting expectations can be useful. Provided that a group of stimuli, corresponding for instance to spectral transitions, feed the same processing unit, the preactivation of this unit by the internal (Context) signal can last a certain period. This allows the set of stimuli to occur within this (possibly large) time interval for them to trigger the unit response. The assignment of stimuli groups to units is performed when the network is enlarged through sprouting (2).

Regarding local variations within a group of stimuli, it is the duration of every signal triggered by stimuli at the unit input which is concerned. In order to detect their coincidence, these signals are delayed by the unit according to the learned reference pattern, and then summed. Summation is efficient only when signals overlap. In order to compensate for differences between the current input and the reference pattern, only rough coincidence is detected between durable signals (see Fig. 5). The more durable the signals, the higher the probability for them to overlap in presence of time distortions.

The duration of signals that are summed by a unit is set when the unit is initialized at sprouting time. Because it is performed in the course of processing, the sprouting of new pathways can be based on the system's expectations. The preactivated parts of memory correspond to the next expected events. If a different configuration of stimuli occurs, it is "stored" in the neighborhood of the expected ones, according to an unsupervised learning process (2). Capability to learn in one shot and without a teacher relies on the fact that what is learnt is not a single vector decision area, but the configuration of many vectors with fixed decision areas. The time duration of this area is a parameter initially provided by the user, and which can be regulated in case of error. In this case, a supervised learning schema is necessary.
5. CONCLUSION

The method proposed in this paper could be considered as an analog and parallel way of treating time distortions, compared to the digital and sequential Dynamic Time Warping algorithm. However, the classical approach is also capable of dealing at the same time with frequency variations induced for instance by the coarticulation effect. Guided Propagation and its associated signal transformations remain to be investigated in this direction, among others.

REFERENCES


(2) Béroule, D.G. The Never-Ending Learning, Neural Computers, R.Eckmiller, C.v.d. Malsburg (Eds), Springer Verlag.


