Neural Expert Model Applied to Phonemes Recognition

Halima Bahi[1], Mokhtar Sellami[1]

[1] LRI laboratory, Computer Science Department, University of Annaba, BP12 23000 Annaba, Algeria
{bahi, sellami@lri-annaba.net}
http://lri-annaba.net

Abstract. Connectionist models often offer good performance in pattern recognition and generalization, and present such qualities as natural learning ability, noise tolerance and graceful degradation. By contrast, symbolic models often present a complementary profile: they offer good performance in reasoning and deduction, and present such qualities as natural symbolic manipulation and explanation abilities. In the context of this paper, we address two limitations of artificial neural networks: the lack of explicit knowledge and the absence of temporal aspect in their implementation. ***STN*** : is a model of a specialized temporal neuron which includes both symbolic and temporal aspects. To illustrate the ***STN*** utility, we consider a system for phoneme recognition.

1 Introduction

The automatic speech recognition (ASR) is the process whereby the machine tries "to decode" the speech signal. Most of the current speech recognition systems are based on hidden Markov models (HMMs) techniques [2],[4]. Another approach besides HMM's are the connectionist techniques[3],[6].

Artificial neural networks (ANNs) are good pattern recognisers, they are able to recognize patterns even when data are noisy, ambiguous or distorted [3]. Albeit, the problem in neural networks is with the choice of architecture (the only way to decide on a certain architecture is on a trial-and-error basis) and the lack of explanation. Researches in this area deal with the integration of symbolic knowledge insight of the connectionist architecture [5],[7].

The purpose of this paper, is to introduce a symbolic neural network dedicated to the speech recognition. The particularity of the proposed system with respect to those available in literature is the introduction of both the temporal and the symbolic aspects.

The remainder of the paper is structured as follows: The second section, defines the automatic speech recognition, then we introduce the ANNs. Section 3, gives an overview of the whole project, which is dedicated to speech recognition. In section 4, we describe the conceptual elements of the first layer. Section 5, describes the

decision layer and particularly the *STN* model. In section 6, practical issues of the application are described for phonemes recognition. Finally, a conclusion is drawn.

2 Speech Recognition and Artificial Neural Networks

2.1 Speech Recognition

The speech recognition task involves several stages, the most important of them is the features extraction stage. In this stage, a smallest set of features will represent the original signal. Then the obtained representation of the signal is compared to the reference patterns to determine the closest one.

Most of the methods used in speech recognition (and in pattern recognition) include two stages : the training and the recognition stages. In the training stage, we present to the system a set of examples and at the end of the stage, the system will be able to distinguish them correctly, in this stage, patterns which were not in the training set are presented to the system, and it should categorize them correctly.

2.2 Artificial Neural Networks

Artificial neural networks are systems composed of a large number of simple interconnected units that simulate brain activity. Each of these units, that are the equivalent of the neuron in a biological simulation, is a part of layered structure, and produces an output that is a non-linear function of the inputs. In the feed forward networks such as the multiplayer perceptron (MLP), the output of each layer of units is connected to units of a higher layer with directed connections that are the equivalent to brain synapses. In the MLP The first layer is called the input layer, the last one is the output layer, and between there may be one or more hidden layers.

2.3 ANNs and Speech Recognition

Figure 1 shows a conceptual block diagram of a speech understanding system loosely based on a model of speech perception in human beings [4]. This diagram clearly underlines the importance of connectionist models when modelling such applications. Early attempts to model speech recognizers uses classical MLP, these approaches assume static representation of the time, later attempts were made to consider the dynamic aspect, the most populer of them is the TDANN (Time Delay Artificial Neural Networks) introduced by Waibel (see [6]), here the time is not explicitly represented in the network and the structure is too much complicated. A connectionist expert system dedicated to speech recognition was presented in [1], although, this system did not consider explicitly the temporal parameter.

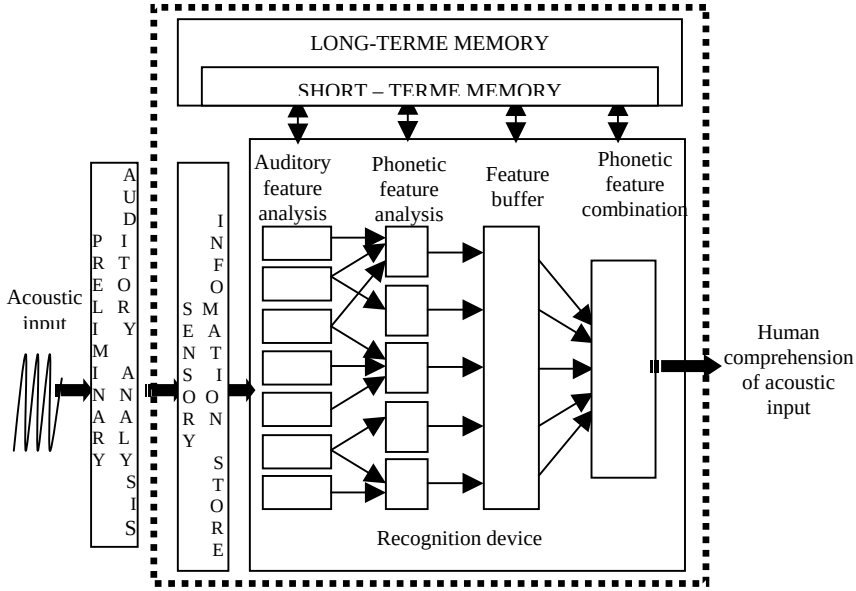

Fig. 1. Conceptual block diagram of human speech understanding (after [4])

3 *NESSR* : Neural Expert System for Speech Recognition

3.1 The System Overview

The overall system comprises three components : a recognition memory, a short term memory and a long term memory. *NESSR* is the recognition memory, which is a neural expert network. *NESSR* is a modular network, the first module is concerned with the phoneme recognition, the second one recognizes the words. The short-term memory is the memory where temporary events are stored, which may occur during the inferencing process (see § 5. 3). The long-term memory is the memory where are stored high level information of the language, this will validate a given decision. The role of this memory is beyond the scope of this paper.

3.2 Integrating symbols insight of the network [1]

We consider an MLP, so neurons are regrouped into layers which correspond to the levels of our application which consists on the isolated word recognition. Thus, the input layer represents the acoustical level, the hidden layer the phonetic level, and the

output layer, stands for the lexical one (figure 2). For the purpose of this paper, we are only concerned with the two first levels. So, the considered objects are : phonemes and their acoustic characteristics.

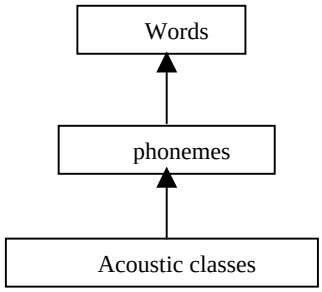

Fig. 2. *NESSR* topology

4 The sensory layer : the acoustic level

The input layer detects changes in the environment. Neurons of the sensory layer captured particularities of the pattern in entry of the network. In the speech recognition context, the cells detect features of the signal.

4.1 The neurons structure : specialized neurons

Since these cells belong to a symbolic network; every cell is specialized in the detection of one characteristic of the signal. We consider these characteristics as acoustic classes, so, we call a neuron of this layer : neuron-class. These particularities did not have a particular physical significance, they are numbered from 1 to n.

4.2 How to determine acoustic classes ?

The first stage of the ASR process provides a collection of numeric vectors from the digitised signal. To translate this representation to a symbolic space, we perform a vector quantization (VQ) over all available vectors in the training stage. VQ enables us to replace each acoustic vector by the correspondent discreet symbol, where symbols represent entries of the code-book. So, there are as many neuron-class as there are entries in the codebook.

4.3 Activation of a neuron

A neuron-class fires if the associated characteristic is detected in the signal. The network dynamic is triggered by discreet instants. At a given instant t, only one neuron-class is active. This supposes that the presentation of a signal to the network

lasts from the instant t_0 to the instant t_n. In this interval of time many neurons can be activated.
We notice that the successive activation of the same neuron is taken into account by the network (see the *STN* model properties).

5 The decision layer : the phonetic level

Activations of the sensory layer are transmitted to the following layer whose role is to associate to the acoustic entries a phonetic units of the language; in this case phonemes. To a detected sequence of acoustic classes will be associated one phoneme. The recognition of a phoneme leads to an implicit segmentation of the signal at this point of the structure.

5.1 Structure of neurons: specialized temporal neurons

As for the sensory neurons the cells of this layer are meaningful. In this case every cell represents one phoneme of the Arabic language, we will call it : neuron-phoneme. A phoneme is defined by the detection of a sequence of acoustic classes. Thus, when there is correlation between the detection of an acoustic class and the recognition of a phoneme, a connection between the concerned neurons is initiated. The activation of these entries must be in a very definite order assured by the structure of the neuron, in which an entry i cannot be considered while the entry i-1 is not already pre–activated. To model such neuron we suggest the following neuron model (figure 3).

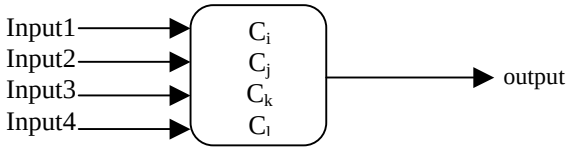

Fig. 3. The *STN* model

5.2 The phoneme characterization

To determine the needed acoustic classes for the detection of a phoneme : we consider the set of classes obtained after the VQ stage, and we operate a study of correlation between these prototypes and the set of phonemes. This permits us to extract the necessary set of classes for the recognition of any phoneme.

Table 1. Line of the correlation matrix

	C1	C2	C3	C4	C5	C6	C7	…	C64
a1_1	×				×	×			

×: detected class

Below is a line of a table illustrating relations between the definite classes and the apparition of a phoneme. This table is automatically built ; each occurrence of a phoneme from the training set is analysed, then quantified. When a characteristic appears in the signal, a mark is set in corresponding phoneme box. If a class appears more than 90% in a phoneme, we consider that it is basic constituent of the phoneme. Once characteristics of the phoneme are designated their order is established.

5.3 Activation of the STN

When a characteristic C_i is detected the associated neuron-class fires and all connections from this neuron are pre-activated. All neuron-phonemes whose first characteristic is C_i are pre-activated. Thus, a neuron-phoneme is pre-activated as soon as its first entry is pre-active, this supposes that several neuron-phonemes can be simultaneously pre-activated. Albeit, a neuron-phoneme fires only if all its entries are activated. When a neuron-phoneme fires all connections coming from the previous layer are deactivated, it is the same way for all competitor neurons, i.e. those which were simultaneously pre-active. If the detection of a characteristic can provoke the activation of more than one target cell, only one cell fires and this information is stored in the short-term memory. Let's notice that this situation is very rare (considering the number of classes 64).

5.4 Illustrative Example

In the figure 8, we present an example to illustrate the particular situations that constitute limit conditions of the model, and justify its use in temporal applications.

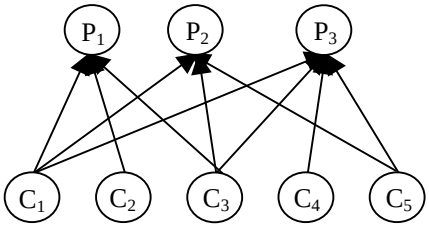

Fig. 4. Network connections to illustrate the *STN* activations

We consider the above network, and we assume the following sequence: …$C_1C_1C_5C_2C_3C_4$…, the activation of the network is transcribed in the following table.

Table 2. Activation example of neuron-phonemes

	Neurone-p1	**Neurone-p2**	**Neurone-p3**
T=0	C1 C2 C3	C1 C3 C5	C1 C5 C4 C3
T=1	C1 C2 C3	C1 C3 C5	C1 C5 C4 C3
T=2	C1 C2 C3	C1 C3 C5	C1 C5 C4 C3
T=3	C1 C2 C3	C1 C3 C5	C1 C5 C4 C3
T=4	C1 C2 C3	C1 C3 C5	C1 C5 C4 C3

Non active Connection
Pre-active Connection

At the instant t = 0, the characteristic C_1 is detected so the correspondent neuron-class fires, and all its output links are pre-activated ; this implies the pre-activation of the three neuron-phonemes of the network; because C_1 corresponds to the first entry for all these target neurons.

At the instant t = 1, the same characteristic is detected, but this second activation doesn't bring any change in the state of the network.

At the instant t = 2, the characteristic C_5 is detected, this induces the pre-activation of the entry C_5 of the neuron-p3. The C_5 entry of neuron-p2 could not be pre-activated, because it could not be considered before the connection C_3 is pre-activated.

At the instant t = 3, the characteristic C_2 is detected this pre-actives the second entry of neuron-p1.

At the instant t = 4, the characteristic C_3 is detected this pre-actives the corresponding entries in the target neurons p1 and p2. In this last pre-activation the neuron-p1 has its entries pre-active. At this moment, it fires and all connections as well as the other target neurons are deactivated.

After C_3 the sequence in entry of the network is segmented and the activation of neuron - p1 is propagated to the following layer. A new session of phoneme recognition starts with C_4.

5.5 Particularities of the STN model

The structure of the STN neuron we suggest to model phonemes, allows the successive detection of the same acoustic characteristic of the signal (C_1 in the previous example); i.e. the model allows stationary transitions of the signal. This structure also allows the insertion of less important classes in the phoneme among pertinent classes (in the previous example C_5 is inserted in p1 structure).

6 Experimental results

To evaluate performances of this module of *NESSR*, we perform some experimentations related to Arabic phoneme recognition. In the following, we describe the practical stages:

6.1 Features extraction

The context of the present work is phoneme recognition. For developing experimental results, a set of Arabic words, including all the phonemes, was used. This corpus comprises twenty five words. Words are segmented and labelled into phonemes. Words are recorded with pause between them, and are uttered by many people of the laboratory. All examples were uttered in relatively quiet room. The incoming signal is sampled at 11025 Hz, with 8 bits of precision, and the sampled signal is processed by a first-order digital filter in order to spectrally flatten the signal. Sections of 400 consecutives samples are blocked into a single frame, corresponding to $400/11.025 \approx$ 36ms. Frames are spaced M samples (M=100). Then the frames are individually multiplied by a N-sample window. In ASR the most-used window shape is the Hamming window. From each frame, we extract a set of 13 Mel Frequency Ceptral Coefficients (MFCCs).

6.2 Vector quantization

We consider all acoustical vectors we obtain during the training stage, we regroup them into disjoint classes (64) using the k-means algorithm. At the recognition phase, the vector quantizer compares each acoustical vector v_j of the signal to stored vectors c_i (code-words), and v_j is coded by the vector c_b that best represents v_j according to some distortion measure d. $d(v_j,c_b) = \min (d(v_j, c_i))$, we use the Euclidian distance.

6.3 Results

The database comprises utterances of 25 words, uttered by 14 speakers, 8 of them participate in the training stage (when, we define the acoustic classes and the phoneme characteristics). To perform evaluation tests we form two groups : The group TS1, includes new utterances of speakers who have participate in the training stage. The group TS2, includes utterances of speakers who did not participate in the training stage and some of those who participate. In the table bellow, we mention results, we have obtained for the considered phonemes (phonemes are given in IPA notation ; /a/, /u/ and /i/ are Arabic vowels).

Table 3. recognition rate in %

	/a/	/u/	/i/	/m/	/ H /
TS1	99.2	98.4	98.2	97.6	95.1
TS2	97.7	97	97	95.8	95

6 Conclusion

In this paper we have attempted to present our contribution in the separate fields of the neurosymbolic systems and the temporal connectionist models. Our suggestion tries to combine in the same network the two components throughout the proposition of a new neuron structure : we called *STN* model. An application of this model is proposed in the phoneme recognition.

Although the obtained recognition rates are under our hope they still being promising ones, and we still believe that the neural expert models are a promising trend in resolution of perception problems, since this category of problems involve both neural models and symbolic reasoning.

References

1. Bahi H., Sellami M., Système expert connexionniste pour la reconnaissance de la parole, proceedings of RFIA, Vol 2, pp: 659-665. Toulouse, France, 2004.
2. Becchitti C., Ricotti L. P., speech recognition*: theory and C++ implementation*, John Wiley, England, 1999.
3. Bishop C. M., Neural networks for pattern recognition , Clarendon Press, Oxford, 1995.
4. Rabiner L., Hwang B., Fundamentals of speech recognition, Prentice Hall, 1993.
5. Sun R., Alexandre F., Connectionist-Symbolic Integration: From Unified to Hybrid Approaches, Lawrence Erlbaum Associates, 1997.
6. Tebelski J., Speech recognition using neural networks, PhD Thesis, Carnegie Mellon University, May 1995.
7. Towell G., Symbolic knowledge and neural networks: Insertion, Refinement and extraction, PhD thesis, University of Wisconsin, Madison, 1991.