A HYBRID HMM/AUTOREGRESSIVE TIME-DELAY NEURAL NETWORK AUTOMATIC SPEECH RECOGNITION SYSTEM

Sid-Ahmed Selouani & Douglas O'Shaughnessy
INRS-Télécommunications, Université du Québec
900 de la Gauchetière Ouest
Montréal, H5A 1C6, Canada
e-mail:{selouani,dougo}@inrs-telecom.uquebec.ca

ABSTRACT

This paper describes a new hybrid approach which aims to significantly improve the performance of Automatic Speech Recognition (ASR) systems when they are confronted with complex phonetic features such as gemination, stress or relevant lengthening of vowels. The underlying idea of this approach consists of dividing the global task of recognition into simple and well-defined sub-tasks and using hearing/perception-based cues. The sub-tasks are assigned to a set of suitable Time-Delay Neural Networks using an autoregressive version of the backpropagation algorithm (AR-TDNN). When they are incorporated in the hybrid structure, the AR-TDNN-based experts act as post-processors of a HMM-based system which thus acquires the ability to overcome failures due to complex language particularities. Results of experiments using either static or dynamic acoustic features show that the proposed HMM/AR-TDNN system outperforms that of the HMM-based system.

1 INTRODUCTION

Architectures of current Automatic Speech Recognition (ASR) systems are generally compact and frontally tackle the global recognition task. The monolithic approach they adopt limits considerably the recognition performance, particularly when they are faced with complex phonetic features and/or prosody-sensitive language [2]. One sees emerging an increasingly marked trend within current research which consists of favoring a scattered architecture of ASR rather than the monolithic one [3]. In this context, we can cite the system described in [6] which is composed of two parts: the first consists of an HMM involved in the recognition of specific phoneme classes and the second is composed of neural networks trained for the disambiguation of pairs such as the /m, n/ nasals. The results showed that significant improvements of ASR scores were obtained for both English and French. To better represent temporal variations in the speech signal, almost all of the mentioned ASR systems add higher-order time derivatives to the set of static parameters.

The approach we propose intends to 'boost' the performance of a modular ASR structure in the case of the complex phonetic features such as gemination, emphasis and relevance of phoneme duration. Our solution consists of placing a hierarchical structure of neural experts (Autoregressive Time-Delay Neural Networks: AR-TDNN) downstream in a baseline HMM-based system. This configuration seems best indicated to exploit the discriminating capacities of neural networks. The final configuration is intended to be more flexible in order to be able to easily generalize the identification of possible new complex features. In order to give additional discriminability for speech pattern comparison, an inclusion of hearing/perception knowledge is carried out through the use of auditory-based cues.

The outline of this paper is as follows. In section 2 we describe autoregressive time-delay neural networks. Next, in section 3 we proceed with a description of the system using the hybrid architecture HMM/AR-TDNN. The hybrid system is evaluated in section 4 by comparing its performances to those obtained by a baseline HMM-based system. In this latter section we also discuss the effect of the use of dynamic auditory-based features.

2 AUTOREGRESSIVE TIME DELAY NEURAL NETWORKS (AR-TDNN)

Because speech is a temporarily unstable phenomenon, we consider Recurrent Networks (RNs) to be more adequate than feedforward networks in the case of any classification task dealing with speech. RNs are generally trickier to work with, but they are theoretically more powerful, having the ability to represent temporal sequences of unbounded length. Another consideration related to phonetic context effects leads us to use a particular RN: the one proposed by Russel [8] and using an Autoregressive (AR) version of the backpropagation algorithm. This type of network can in principle capture naturally the coarticulation phenomenon of speech. However, even if RNs using AR perform very well in the context-dependent labelling, this power turns out to be

source of a disappointment in the case of phoneme time shifting. The approach we are investigating proposes to integrate, in addition to the AR component, a delay component similar to the one used by Waibel's Time-Delay Neural Networks (TDNN) [9]. Through this combination, we expect that the ability of the system to discern the phonological length even in a strong coarticulation context will be increased. The model described by Russel [8] includes an autoregressive memory which constitutes a form of self-feedback where the output depends on the current output plus a weighted sum of previous outputs. Then, the classical AR node equation is:

$$y_i(t) = f\big(bias + \sum_{j=1}^{P} w_{i,j}\, x_j(t)\big) + \sum_{n=1}^{M} a_{i,n}\, y_i(t-n), \quad (1)$$

where $y_i(t)$ is the output of node i at time t, *f(x)* is the *tanh(x)* bipolar activation function, P is the number of input units, and M is the order of autoregressive prediction. Weights $w_{i,j}$, biases, and AR coefficients $a_{i,n}$ are adaptive and are optimized in order to minimize the output error. Our proposition consists of incorporating a time delay component on the input nodes of each layer and then Equation 1 becomes:

$$y_i(t) = f\big(bias + \sum_{m=0}^{L}\sum_{j=1}^{P} w_{i,j,m}\, x_j(t-m)\big) + \sum_{n=1}^{M} a_{i,n} y_i(t-n), \quad (2)$$

where L is the delay order at the input. Feedforward and feedback weights were initialized from a uniform distribution in the range $[-0.8, 0.8]$. A neuron of the AR-TDNN configuration is shown in Figure 1. An autoregressive backpropagation learning algorithm performs the optimization of feedback coefficients in order to minimize the mean squared error noted *E(t)* and defined as:

$$E(t) = \frac{1}{2}\sum_{i}\big(d_i(t) - y_i(t)\big)^2, \quad (3)$$

where d_i is the desired value of the i^{th} output node. The weight and feedback coefficient changes, noted respectively $w_{j,i,m}$ and $a_{i,n}$, are accumulated within an update interval $[T_0, T_1]$. In the proposed AR-TDNN version, the update interval $[T_0, T_1]$ is fixed such as it corresponds to the time delay of the inputs. The updated feedback coefficients are written as follows:

$$a_{i,n}^{new} = a_{i,n}^{old} + \frac{1}{T_1 - T_0}\sum_{t=T_0}^{T_1} \Delta a_{i,n}(t), \quad (4)$$

and if T is the frame duration, the weights are as follows:

$$w_{i,j}^{new} = w_{i,j}^{old} + \frac{1}{LT}\sum_{t=T_0}^{T_1} \Delta w_{i,j}(t). \quad (5)$$

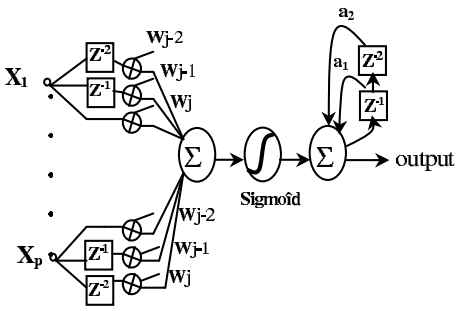

Figure 1: *AR-TDNN unit.*

The calculation of $\Delta a_{i,n}(t)$ variation is detailed in [8]. The optimization of weights and biases are performed as in Waibel's networks [9]. Hence, the $\Delta w_{i,j}$ variations are accumulated during the update interval after accumulating Time-Delay frames at the input.

2.1 Ear-Based Acoustic Attributes

Cues derived from hearing phenomena studies are extracted thanks to the Caelen ear-model [1]. In this model, the internal ear is represented by a coupled filter bank where each filter is centred on a specific frequency. The filters' number can be limited to 24 covering a 16 Hz-12000 Hz frequency range. The 24-channel spectrum obtained in the output of the 24 coupled filters can be used directly as input data. Furthermore, from a particular linear combination of the outputs of these channels, 7 cues are derived: acute/grave (AG), open/closed (OC), diffuse/compact (DC), sharp/flat (SF), mat/strident (MS), continuous/discontinuous (CD) and tense/lax (TL). These indicative features are very relevant to characterize the phonemes of many languages [5]. Over each phone, an average of the ear-based indicative features is calculated and used as inputs in the AR-TDNN experts.

2.2 AR-TDNN vs. TDNN

AR-TDNNs are trained by using the Nguyen-Widrow initialization conditions [7]. The TDNN part of the system consists of three layers. Each unit in the hidden layer receives input from the coefficients in the three-frame window of the input layer. The input is centred around the hand-labelled phones. Experiments using approximately 6000 phonemes uttered by six speakers are carried out in order to compare the performances of AR-TDNNs and TDNNs. The results given in Figure 2 show that the AR-TDNN with a macro-class recognition rate average of 82% surpasses significantly the standard backpropagation-based system with a global $77,5\%$. A significant difference of accuracy in favour of AR-TDNN is observed in the case of semantically-relevant lengthening of phonemes. This experiment confirms the capability of the AR-TDNN configuration to simultaneously perform context-sensitive decisions and to capture the temporal component.

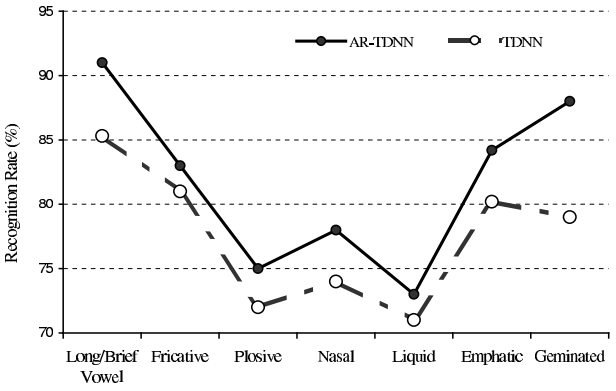

Figure 2: *Comparison of TDNN and AR-TDNN performances over macro-classes.*

3 HMM/AR-TDNN HYBRID STRUCTRE

Training the HMM/AR-TDNN on an utterance proceeds in two steps. The first step performs optimal alignment between the acoustic models of phones and the speech signal. In the second step the AR-TDNN system acts as post-processor to the HMM-based system and refines its recognition results. The global task is then divided between the main system constituted by the HMM-based system and the 'booster' system composed of AR-TDNN. We require HMM to achieve phone identification without discriminating between long and short vowels and between emphatic and non-emphatic consonants. The gemination detection is also not required. The hand-labelled data set input to HMM presents a single label for phonemes belonging to these macro-classes. For instance, in the case of short vowel /a/ and long vowel /a:/, a unique /A/ label is given. The /A/ sequence of phones is presented to the AR-TDNN system which makes final and finer decisions related to the long/short vowel discrimination.

Because of the importance of the phonetic context for performing phoneme identification, a careful analysis must be done for selecting the learning set. The supervision of this learning considers phones as complete items. The coarticulation effect makes this supervision difficult. The adopted solution consists of executing the learning phase as if a phone of the target-phoneme appears in the speech continuum, the AR-TDNN activation arises gradually in the output. In the example of emphatic consonant detection/classification, the task consists of learning to recognize the following sequence: LCE-EMPH-RCE: LCE is the left phonetic context of the emphatic consonant (noted LCE) and RCE is its right phonetic context. *EMPHA_NET* (emphasis expert network) receives three input tokens at a time t and it must detect an emphatic sequence from any other sequence combination. The learning sets the output at the high level $(+1)$ when the end of the LCE-EMPH-RCE sequence is attained. The low level (-1) is set otherwise, i.e. if a scrolling (stream) of non-emphatic

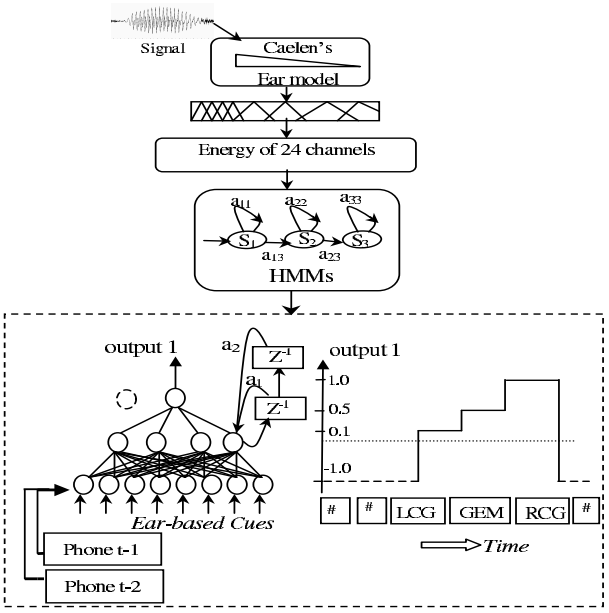

Figure 3: *General overview of the hybrid ASR system.*

phone sequences is observed. An autoregressive order of 2 is chosen and a delay of 2 phones is also fixed. These lower values of delay and order are justified by the fact that phones are used instead of frames. Consequently the stability of AR nodes is ensured. Besides the *EMPHA_NET* system, other AR-TDNN-based expert systems are provided: *DURA_NET* and *GEMI_NET*. They respectively perform long-short vowel discrimination and Geminated-Non-Geminated opposition detection. These tasks are accomplished according to the same protocol used by *EMPHA_NET*.

4 EXPERIMENTAL RESULTS

In order to recognize the speech data, the HTK-based speech recognition system described in [4] has been used throughout all experiments. HTK is an HMM-based speech recognition system. The toolkit can be used for isolated or continuous whole-word/phone-based recognition systems. The toolkit was designed to support continuous-density HMMs with any number of state and mixture components. In all our experiments, 24 coefficients represent the outputs of the filter bank which simulate the basilar membrane of the ear. These coefficients were calculated on a 30-msec Hamming window advanced by 10 msec each frame. This vector constitutes a 24-dimensional static vector upon which the HMMs, that model the speech subword units, were trained. The baseline system for the recognition task uses mono-Gaussian mixture HMM system. As is mentioned in section 2.1, 7 ear-based indicative features are used as AR-TDNN input. This vector is expanded by a component representing the middle ear energy [1]. Thus, in a first experiment an 8-dimensional (static) vector is used by the AR-TDNN experts. In a second experiment,

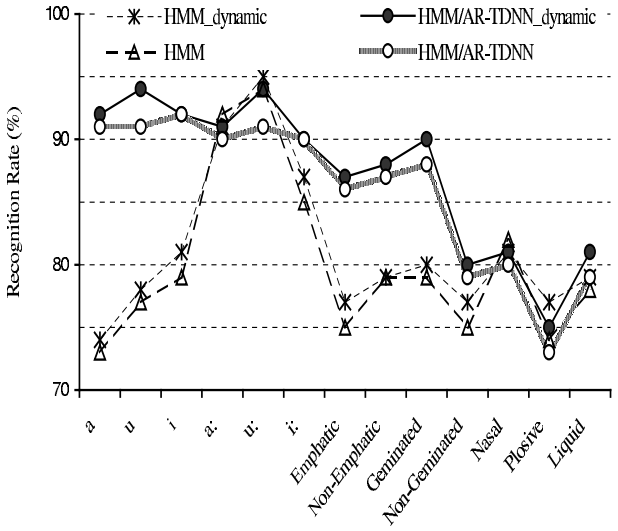

Figure 4: *Percent phoneme recognition performance of HMM and HMM/AR-TDNN using static and dynamic parameters.*

dynamic parameters are considered. The first derivatives of the 8 components of the AR-TDNN input are embedded in the original vector in order to constitute a 16-dimensional (static+dynamic) vector. In the HMM part, a decimation is used to constitute a 12-dimensional static vector, which is expanded by its first derivative to produce a 24-dimensional (static+dynamic) vector.

We compare the hybrid HMM/AR-TDNN system to a baseline HMM-based system. These results concern 60 VCV utterances and 50 phrases pronounced by 6 speakers. As a whole, the test concerns 3724 vowels (1348 long), 1197 fricatives (182 geminated, 193 emphatics), 1089 plosives (215 geminated, 273 emphatics), 573 nasals and 413 liquids. The analysis of the results revealed that the hybrid configuration is more accurate in all cases of complex phonemes. In the case of static analysis, we found that the HMM/AR-TDNN system achieved 86% accuracy, which represents 6% fewer errors than the HMM baseline system. Concerning the standard HMM, we noticed that it failed dramatically in the discrimination of long and brief vowels. An imbalance of performances reaching 20% in the case of /a:/ and /a/ vowels is observed. In the case of emphatic consonants, the hybrid system performs with 12% fewer errors than standard HMM. The same trend is observed when dynamic parameters are used. The hybrid and monolithic systems using dynamic features obtained, respectively, 87% and 81% phoneme recognition rates. We must underline the fact that the improvement reached by the structural modification of the ASR system is more significant than the inclusion of dynamic features : 6% vs. 1%. The cumulative improvement is about 8% if both hybridizing and inclusion of first derivatives are considered.

5 CONCLUSION

A hybrid approach for speech recognition was presented. Our objective was to test the ability of a system combining HMM and AR-TDNN to detect features as subtle as gemination, emphasis and relevant lengthening of vowels. This hybrid system has been compared to a baseline HMM-based system. Considering the obtained results, it seems clear that the proposed hybrid HMM/AR-TDNN approach improves significantly (8%) performances of standard HMM. The split of the global speech recognition task into subtasks assigned to more adapted systems, conjugated with the use of dynamic ear-based features, constitutes from our point of view a powerful and promising way to overcome problems due to language particularities. Hence, the generalization can easily be considered in the context of multi-lingual ASR.

6 REFERENCES

[1] J.Caelen, "Space/Time Data-Information in the ARIAL Project Ear Model", Speech Communication, Vol. 4, pp. 163-179, 1985.

[2] R. Cole et al, "The Challenge of Spoken Language Systems: Research Directions for the Nineties", IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, pp. 1-20, 1995.

[3] G.D. Cook, S.R. Waterhouse, and A.J. Robinson, "Ensemble Methods for Connectionist Acoustic Modeling", Proc. EUROSPEECH'97, pp. 1959-1962, 1997.

[4] Cambridge University Speech Group, "The HTK Book (Version 2.1.1)", Cambridge University Group, March 1997.

[5] R. Jakobson, G. Fant, and M. Halle, "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates", MIT Press, Cambridge, 1963.

[6] J.F. Mari, "HMM and Selectively Neural Networks for Connected Confusable Word Recognition", International Conference Speech and Language Processing (ICSLP), pp. 1519-1522, 1994.

[7] D. Nguyen, and B. Widrow, "Improving the Learning Speed of Two-Layer Neural Networks by Choosing Initial Values of the Adaptative Weights", Proc. of IJCNN (III), pp. 21-26, 1990.

[8] R.L. Russel, and C. Bartley "The Autoregressive Backpropagation Algorithm", Proc. of IJCNN (II), pp. 369-377, 1991.

[9] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. on Audio Speech and Signal Processing, No. 37, pp. 328-339, 1989.