Continuous Speech Recognition Using a Context Sensitive ANN and HMM2s

Nicolas Pican, Jean-François Mari, D. Fohr

pican@loria.fr, jfmari@loria.fr, fohr@loria.fr

CRIN-CNRS & INRIA Lorraine
BP 239, F-54506 VANDŒUVRE-les-NANCY Cedex, FRANCE
Tel: (33) 83.59.20.54, Fax: (33) 83.41.30.79

ABSTRACT

The phonetic context has a large effect on phonemes in a continuous speech signal [1]. Therefore recognition systems that model allophones using context-dependent Hidden Markov Models have been implemented [4]. Second-order HMMs (HMM2s have a great ability for the segmentation in the temporal domain [6][7] but have some difficulties in the recognition because the MLE training (Maximum Likelihood Estimation) is not discriminant, whereas the discrimination is one of the abilities of the Artificial Neural Networks models. In the last three years we have developed a new ANN model named OWE (Orthogonal Weight Estimator)[10][11].

The principle of the OWE is a ANN that classifies an input pattern according to contextual environment. This new ANN architecture tackles the problem of context dependent behaviour training. Roughly, the principle is based on main MLP (Multilayered Perceptron) in which each synaptic weight connection value is estimated by another MLP (an OWE) with respect to context representation. In this paper, we present 2 hybrid systems for phoneme recognition. In both systems, 48 context independent HMM2s segment the input signal. In the first system, the OWE performs the labelling of segments and, in the second system, the OWE outputs are the input frames of the HMM2s. Experiments on TIMIT range from 56% to 67% accuracies on the 48 phonemes set.

1. INTRODUCTION

This paper addresses the problem of continuous speech recognition using an hybrid approach based on stochastic modeling with hidden Markov models (HMM2s) and artificial neural networks (ANN). One of the main problem in phonemes recognition is the modeling of the contextual variations. For example, the coarticulation effects in continuous speech causes vowel formants tracks to be affected by nearby phonemes, or stop burst modified by the following phonemes [1]. Therefore, many recognition systems that model allophones using context-dependent Hidden Markov Model have been implemented [4][5].

HMM2s have a great capability for performing the segmentation in the temporal domain [6]. In most cases, the HMM2s are trained using the maximum likelihood estimation (MLE) paradigm. This leads to a non optimal capability in discriminating temporal segments in the recognition process because the models are not competitive in the training process. In the other hand, the ANN are discriminative but experience some difficulties to perform a temporal segmentation.

In the last three years, we have developed a new ANN model named OWE (Orthogonal Weight Estimator)[10][11]. This new ANN architecture tackles the problem of the context dependent behavior training. Merely, the principle is based on main MLP (Multilayered Perceptron) in which each connection is estimated by one MLP (an OWE) fed by a context representation. HMM2s and OWEs have already been used in an hybrid system [13] for recognizing the English stop consonants (/p,t,k,b,d,g/) of the TIMIT database. The results validate this approach by showing an increase in the recognition accuracy compared with a pure HMM2 system.

This paper presents an extension of this hybrid system to the recognition of the 48 English phonemes in continuous speech using the TIMIT speaker-independent database.

The paper is organized as follow: the section 2 gives a short description of HMM2s. Section 3 describes the OWE ANN. In section 4, we give results on the TIMIT database and discuss them.

2. HMM2 FRAMEWORK

In a second-order HMM, the underlying state sequence is a second-order Markov chain in which the probability of transition between two states at time *t* depends on the states in which the process was at time *t-1* and *t-2*. The output state probability is represented by a mixture of gaussian estimates with full covariance matrices.

Notations

We call:

- λ , the second-order hidden Markov model,
- b_i the density associated to state i,
- O_t observation at time t (dimension D),
- $P(O/\lambda)$ the likelihood of the sequence of observations $O_1,O_2...O_T$ assuming model λ ,
- $\aleph(\mu, \Sigma)$ the normal probability density function (pdf) of dimension *D* with mean μ and covariance matrix Σ .

Increasing HMM order

Usually, the transition probabilities of HHM1 are:

$$P(S_t = k|_{S_{t-1} = j,\, S_{t-2} = i,\, S_{t-3} = \ldots}) = P(S_t = k|_{S_{t-1} = j}) = a_{jk}$$

Many researchers have noticed that these probabilities have a negligible impact on the recognition rate and are often ignored. In HMM2 they become:

$$P(S_t=k|_{S_{t-1}=j,\, S_{t-2} = i,\, S_{t-3} = \ldots}) = P(S_t=k|_{S_{t-1}=j,\, S_{t-2}=i}) = a_{ijk}$$

The pdf associated to state s_i and the likelihood of vector x given $\aleph(\mu, \Sigma)$ can be expressed by:

$$b_i(O_t) = \sum_m c_{im} \aleph(\mu_m, \Sigma_m, O_t) \qquad \sum_m c_{im} = 1$$

$$\aleph(\mu, \Sigma, x) = \frac{1}{\sqrt{(2\pi)^D det\ \Sigma}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

The generation of "Forward-Backward" functions are obtained by adding an index indicating where the process was at time *t-2*.

$$\alpha_{t+1}(j,k) = \sum_{i=1}^{N} \alpha_t(i,j) a_{ijk} b_k(O_{t+1}), \quad 2 \le t \le T-1$$

$$\beta_t(i,j) = \sum_{k=1}^{N} \beta_{t+1}(j,k) a_{ijk} b_k(O_{t+1}), \quad 2 \le t \le T-1$$

The count associated with transition *(i, j, k)* becomes:

$$\eta_t(i, j, k) = \frac{\alpha_t(j,k) a_{ijk} b_k(O_{t+1}) \beta_{t+1}(j,k)}{P(O/\lambda)}, \ 1 \le t \le T-1$$

A more extensive presentation can be found in [6][7].

Using these definitions, the maximum likelihood estimation is straightforward [6].

3. OWE FRAMEWORK

We propose in this section the presentation of the main principles of the contextual ANN, named OWE (Orthogonal Weight Estimator).

3.1. Introduction

One of the better known and used ANN architecture in classification problem is indisputably the multilayered Perceptron (MLP) [3]. Even if the results obtained with this architecture are the best in an unvarying contextual environment, they become very poor when the perceptions about an object, that has to be classified, change with respect to the variation of the context.

Based on the result that a weight value of a connection in a MLP changes continuously with respect to a continuous variation of a context parameter [9], we have define a contextual ANN architecture in which each synaptic weight value of a MLP is computed by an OWE (another MLP) fed by the contextual parameters.

3.2. Connectionism point of view

The main usual connection type in MLP models is the axo-dendritic connection. This connection type is based on the fact that the axon of an afferent neuron is connected to another neuron via a synapse on a dentrite (Figure 1)

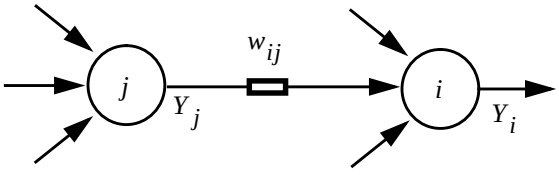

Figure 1: classical connection type

The formalization of the relaxation phase of one neuron *i* in a classical MLP architecture reads $Y_i = F_i(\sum_J w_{ij} Y_j)$, where Y_i and Y_j are respectively the post synaptic activity of neuron *i* and neuron *j*, w_{ij} is the synaptic efficiency of the connection between neuron *j* and neuron *i*, $J = \{j_1, ..., j_n\}$ is the set of afferent connections of the neuron *i*, and $F_i()$ is the transfer function of neuron *i* (usually a sigmoid function).

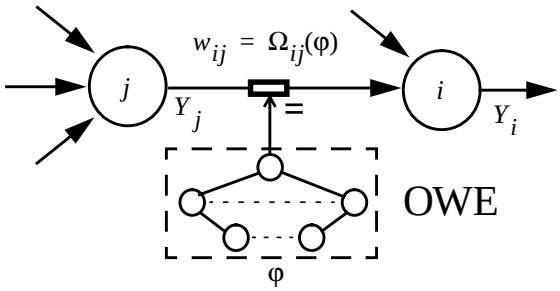

Figure 2: OWE connection type

The principle of the OWE is a ANN that classifies an input pattern *x* according to contextual environment φ.

An OWE architecture, defined by the connection type (Figure 2), is a main MLP and a set of other MLPs, called the OWEs. Each OWE is used to compute the efficiency of each synapse *ij* in the main MLP. Thus the post-synaptic activity of a neuron *i* in the main MLP becomes $Y_i = F_i(\sum_J \Omega_{ij}(\varphi) Y_j)$ where $\Omega_{ij}(\varphi)$ is the weight value function of the connection *ij* with respect to a contextual parameter φ which is approximated by a MLP, the OWE neural network**.**

The training algorithm for this architecture consists to use for each pattern (x, φ) the gradient of the error of each connection in the main MLP, classically computed by a backpropagation algorithm, as the output error of each OWE. Thus these output error signal are used to train each OWE to compute $\Omega_{ij}(\varphi)$**.** This algorithm called “An On-line Learning Algorithm for the Orthogonal Weight Estimation of MLP” is fully detailed in [10].

3.3. Internal structure of OWE

We use a 24x24x48[1] local feedforward MLP with a bias for the main MLP, and a 48x8x1[2] local feedforward MLP with bias for each OWE (Figure 3). The main MLP is fed by the static and dynamic coefficients of the current frame, denoted as B in Figure 3. Each of the 1800 OWEs is fed by the static and dynamic acoustic coefficients of the left context, denoted as A, and the right context, denoted as C.

A parallel implementation of this OWE architecture have must be done on an Intel Paragon parallel computer with 56 nodes and a Silicon Power Challenge Array using MPI (Message Passing Interface) development tools [12].

---

1. 24 input neurons, 24 neurons in the hidden layer, 48 output neurons (one for each phoneme)
2. 48 input neurons, 8 neurons in the hidden layer, 1 output neuron (for the value of weight in the main MLP)

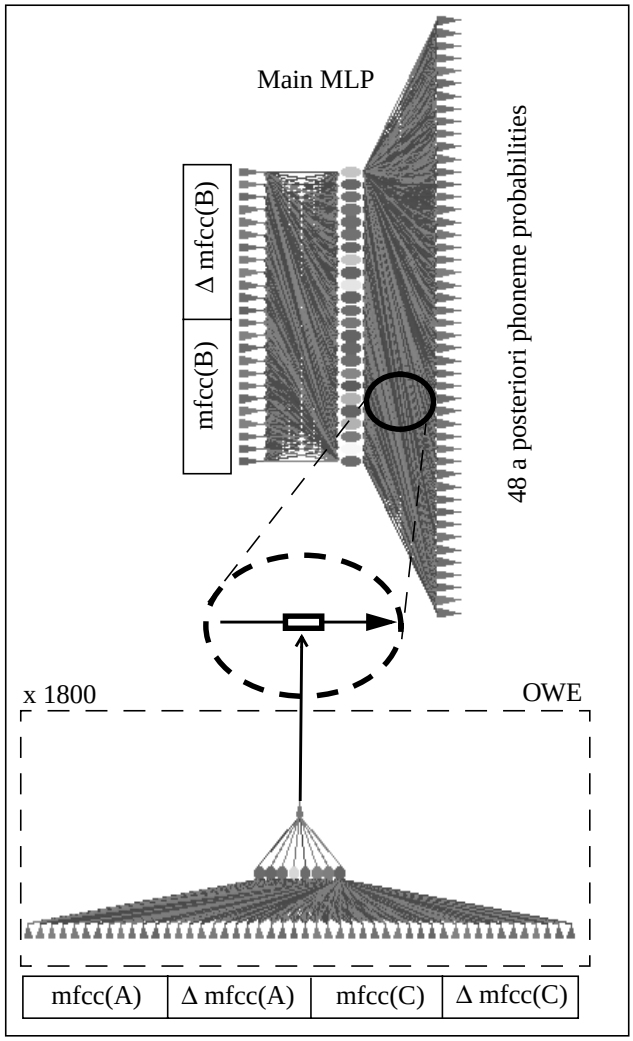

Figure 3: the OWE Architecture recognizer

4. TEST PROTOCOL AND RESULTS

4.1. Database

To further assess the modeling capabilities of HMM2 plus OWE, we developed 2 hybrid phone recognizers using the TIMIT database. During the recognition experiments, a phone-based bigram is used. The results are given on the set of 39 phonemes as defined in [5]. For the experiments, we used the training/test subdivision as specified by the TIMIT-CDROM:

- training set: 8 sentences spoken by 462 speakers,
- test set: 8 sentences spoken by 168 speakers,

We also excluded the "sa" sentences from the training and testing sets.

4.2. Acoustic analysis

For the speech representation, we compute 12 static MFCC coefficients on a 32 ms window every 10 ms. We also concatenate 12 first-order regression coefficients and 12 second-order regression coefficients to the static ones.

Each A,B and C frame is constituted by the last 11 static coefficients (we remove the first coefficient C_0, called loudness) plus the 12 first-order regression coefficients plus the first second order regression coefficient $\Delta\Delta C_0$. We only use the B frame in the HMM2.

A and C frames are taken respectively 5 frames before the B frame (-50 ms) and 5 frames after the B frame (+50 ms).

4.3. Training

48 Context-independent phoneme HMM2 are trained using MLE paradigm on the whole training set. In parallel, OWE architecture is trained using "On-line Learning" algorithm. The same current frame (B) is used for both models.

4.4. Testing

To compare the potentiality of each parts of the 2 hybrid systems, we did 5 experiments.:

• Frame labeling using only OWE architecture (OWE / F)

In this part our interest is to underline the capacities of OWE in the frame labelling task. Using a winner-takes-all paradigm we label each presented frame C according to is context (A,C) by taken the winner phoneme corresponding to the phoneme (output node) that have the maximum *a posteriori* probability. The frame accuracy reaches 66.0 %.

• Hand segmentation & OWE recognition (h-OWE)

Here, we use the hand segmentation given in the Timit database. The label is the *argmax* of the product of the *a posteriori* probabilities of each frame belonging to the segment. The phoneme accuracy reaches 72.3 %.

• HMM2s segmentation & OWE recognition (H+OWE)

In this experiment, the 48 HHM2s perform a segmentation of the utterance but the labelling is done by the OWE as in the *h-OWE* experiment. The phoneme accuracy was disappointing and floors at 56%. Our main explanation is the bias introduced by the hand segmentation that was used to train the OWE as already mentioned in [13]

• OWE as preprocessing of HMM2s (OWE+H)

Here, we use both models is completely different way. The HMM2's frames are no longer the frames B (35 static and dynamic coefficients) but are the 48 outputs of the OWE architecture.

Each HMM2 is a 3 states, left-right, self loop HMM2. The 3 states are tied to the same pdf. We tried different kinds of *pdf*. First, we have estimated the mean and covariance of the frames that were associated to a state during the training as they followed a normal law. But, we noticed that the covariance matrices were singular (some diagonal terms were too low). We next, used an identity covariance matrix and so computed the quadratic distortion between the mean and the input frame. We also tried the kullback-leibler distance but the best performances were obtained by using the log of the a posteriori probability of the phoneme given the

acoustic (corresponding to the component of the 48 OWE outputs). In this case, the accuracy reaches 67%.

The following table summarizes the recognition rates on the 5 experiments. We use the system with pure HMM2s as a reference system.

	OWE / F	HMM2	h-OWE	H+OWE	OWE+H
Accuracy	66.4	70.6	72.3	56.0	67.0

Table 1: Recognition rates on 39 phonemes set

5. CONCLUSION

We have presented 2 hybrid recognition systems based on HMM2 and OWE. Even if the reference system gives the best results (HMM2), we have shown the great capabilities of OWE architecture on context-dependent classification (OWE / F and h-OWE). compared to other works [14].

Our ongoing work concerns the improvement of the discriminating power of the new ANN-OWE that will increase the performances of OWE+H hybrid system. An other major issue is to accurately model the states output pdf in the OWE+H architecture. We also work on a signal segmentation method using the variations of OWE outputs.

Aknowledgements: we thank IRISA (computer science research institute at Rennes, France) and LIP (parallel computer science research laboratory at Lyon, France) for the access to their Paragon Intel parallel computers. We also thank CCH (Charles Hermite Center at Nancy, France) for the access to their SGI Power Challenge Array.

6. REFERENCES

1. Bonneau, A., Coste-Marquis, S., Laprie, Y. "Strong Cues for Identifying Well-Realized Phonetic Features". In Proceedings of International Congress of Phonetic Science, pp 144--147. Stockholm (Sweden), nov 1995.
2. H. Bourlard and N. Morgan: Connectionist Speech Recognition: a Hybrid Approach, Kluwer academic publishers, 1994
3. Hertz, J., Krogh, A., Palmer, R. "Introduction to the Theory Of Neural Computation. Santa Fe Institute. Addison Wesley Ed., Lecture Notes Vol I. March 1992
4. Lamel, L.F., Gauvain, J.L. "High Performance Speaker-Independent Phone Recognition Using CDHMM". In proceedings of EuroSpeech, Berlin. Vol. 1, pp 121-124. September 21-23, 1993.
5. Lee, K.F., and Hon, H.W “Speaker-Independent Phone Recognition Using Hidden Marken Models”, IEEE Trans. ASSP, 37 (11), 1989
6. Mari, J.F, A. Kriouile and Haton, J.P "Automatic Word Recognition Based On Second-Order Hidden Markov Models" IEEE Trans. on S.A.P., Vol 1, no. 5, pp 22-25, Jan. 1997.
7. Mari, J.F., Fohr, D., Junqua, J.C. "A Second-order HMM for High Performance Word and Phoneme-based Continuous Speech Recognition. In IEEE ICASSP Atlanta 1996.
8. Mari, J.F., Fohr, D., Anglade, Y., and Junqua, J.C. "Hidden Markov Models and Selectively Trained Neural Networks for Connected Confusable Word Recognition". ICSLP, pp S26-11. Yokohama, Japan. 1994.
9. Pican, N., Fort, J.C, Alexandre, F. "A Lateral Contribution Learning Algorithm for multi MLP Architecture". ESANN Proceedings. D Facto, Brussels, 20-22 April 1994.
10. Pican, N., Fort, J.C, Alexandre, F. "An On-line Learning Algorithm for the Orthogonal Weight Estimation of MLP". Neural Processing Letters, D facto, Brussels, Vol 1, No 1, pp 21-24, 1994.
11. Pican, N., Alexandre, F.: "How OWE Architectures encode Contextual Effects in ANNs". Mathematics and Computers in Simulation. 41 (5-6) July 1996.
12. Pican, N. :"Intrinsect and Parallel performances of the OWE Neural Network Architecture". ICANN proceedings. Bochum, Germany. 17-19 July 1996.
13. Pican, N., Fohr, D., Mari, J.F.: "HMMs and OWE Neural Network for Continuous Speech Recognition". International Conference on Spoken Language Processing, ICSLP 96 Philadelphia, 1996.
14. Riis, S. K., Krogh, A. : "Hidden Neural Networks: A Framework for HMM/NN Hybrids". ICASSP’97, pp 3233-3236, Munich, 1997.