HMMs and OWE Neural Network for Continuous Speech Recognition

Nicolas Pican, Dominique Fohr, Jean-François Mari

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

CRIN-CNRS & INRIA Lorraine
BP 239, F-54506 VANDŒUVRE-les-NANCY Cedex, FRANCE
Tel: (33) 83.59.20.53, Fax: (33) 83.41.30.79

ABSTRACT

The phonetic context has a large effect on stop consonants in a continuous speech signal [1]. Therefore recognition systems that model allophones using context-dependent Hidden Markov Models have been implemented [3]. HMMs have a great ability for the segmentation in the temporal domain [4][6] 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)[9][10].

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 a hierarchical system for phoneme recognition: first the system segments the input signal using 48 context independent HMMs. Then the stop consonant are reordered by a OWE ANN. Experiments on TIMIT show 78 % of correct recognition rate on the 6 stop consonants (/p, t, k, b, d, g).

1. INTRODUCTION

This paper addresses the problem of stop consonant recognition in continuous speech. One major difficulty is to model properly the influence of the phonetic context. Various studies have been conducted in the framework; some of them have yield to the specification of context dependent HMMs [3], other have investigated hybrid system based on MPL plus HMM [7], or AI technics [1].

The standard way to integrate the context parameters is to specify a grand vector made up with the input pattern and its contextual frames. But recent works [8][9][10] show that performances are higher when contextual parameters are treated separately like in a new architecture called OWE that may be viewed as a MLP where the weights are function of the contextual parameters.

In this architecture, the main MLP is fed by the central frame of the burst. The context is represented by two central frames: one for the closure and one for the following segment. Both central frames fed each OWE. The segmentation in phonemes is carried out by HMMs.

The recognition is a hierarchical process where the HMM gives a segmentation and the final classification is done by the OWE.

The paper is organized as follow: the section 2 gives a short description of the second order HMMs. Section 3 describes the OWE ANN. In section 4, we give results on the TIMIT database and discuss them in section 5.

2. HMM FRAMEWORK

In a second-order HMM (HMM2), 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:

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

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

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:

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

A more extensive presentation can be found in [5][6].

Using these definitions, the maximum likelihood estimation is straightforward [5].

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) [2]. 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 must to be classify, 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 [8], 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 parameter.

3.2. Connectionnism 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)

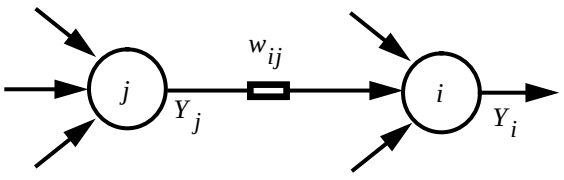

Figure 1: classical connection type

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).

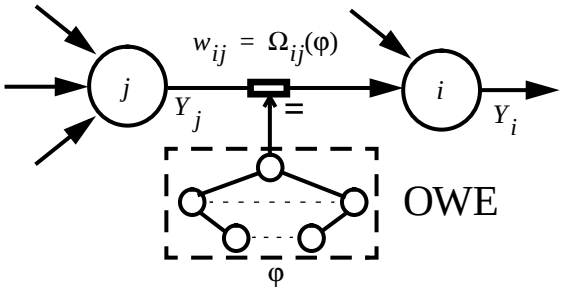

Figure 2: OWE connection type

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 [9].

3.3. Internal structure of OWE

We use a 22x12x6[1] local feedforward MLP with a bias for the main MLP, and a 33x6x1[2] local feedforward MLP with bias for each OWE (Figure 3). The main MLP is fed by the static and dynamic coefficients of the central frame of the burst, denoted as B in Figure 3. Each of the 354 OWEs is fed by the static and dynamic acoustic coefficients of the central frame of the closure, denoted as A in Figure 3, plus 11 static MFCC of the central frame of the following segment, denoted as C in Figure 3.

A parallel implementation of this OWE architecture have been done on an Intel Paragon parallel computer with 56 nodes [11].

1.22 input neurons, 12 neurons in the hidden layer, 6 output neurons (one for each stop consonant)

2.33 input neurons, 6 neurons in the hidden layer, 1 output neuron (for the value of weight in the main MLP)

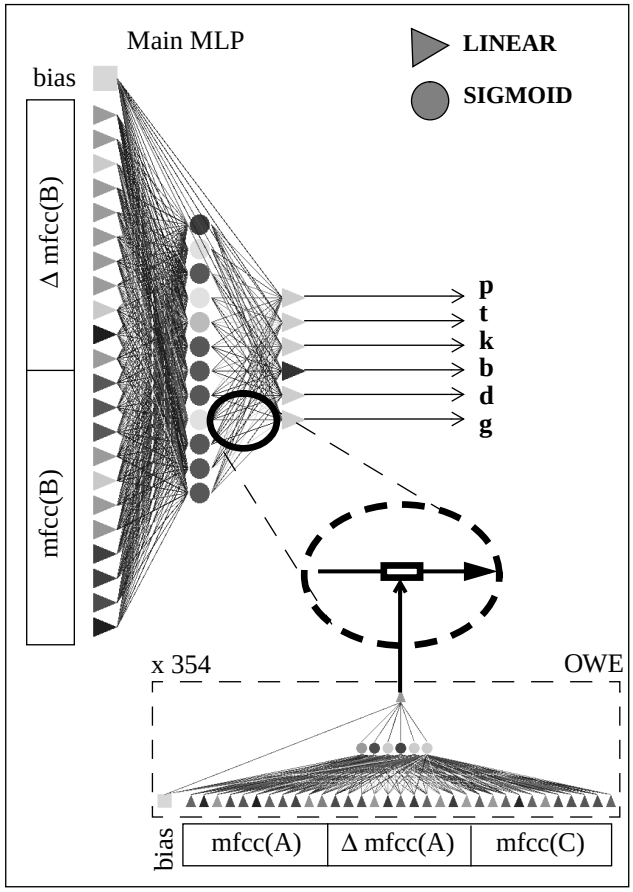

Figure 3: the OWE Architecture recognizer

4. TEST PROTOCOL AND RESULTS

4.1. Database

To further assess the modeling capabilities of HMM2 plus OWE, we developed a phone recognizer using the TIMIT database. During the recognition experiments, a phone-based bigram was used. To compute the results, the set of 39 phonemes as defined in [7] was used. 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 to the static ones.

In the HMM feature vector used for segmentation purposes, we remove the first coefficient, C_0, called loudness but we use instead the second order regression coefficient $\Delta\Delta C_0$.

In the OWE feature vector used for recognition purposes, we remove both C_0 and $\Delta\Delta C_0$.

4.3. Training

Context-independent phoneme HMM models are first trained using MLE paradigm on the whole training set. Then the training subset is labelled using these HMM models constrained to the manual labelling given in the TIMIT CDROM.

In order to train the OWE recognizer, we have to extract the central frames of the three consecutive segments (A,B,C).

Two training experiments have been conducted depending of how the signal has been segmented. In the first one, we train the OWE architecture on the TIMIT hand labels. In the second one, we use the HMM segmentation.

4.4. Testing

4.4.1 Using forced HMM segmentation

Each sentence drawn from the test set is first segmented using the HMM models. At this point, we do not recognize the phonemes, but we look only for the borders of the segments; thus, there are neither substitution, insertion, nor deletion errors. According to this segmentation, for each stop segment (/p, t, k, b, d, g/) we extract the three central frames (A, B and C), and feed the OWE with them. The winner-takes-all answer is compared to the hand labelled segment.

	p	t	k	b	d	g
p	676	53	45	64	14	1
t	60	1084	97	9	119	4
k	42	69	899	4	16	62
b	84	13	4	677	39	18
d	14	94	15	28	491	25
g	11	10	76	16	56	301
nb of uttered phonemes	887	1323	1136	798	735	411
% correct	76%	82%	79%	85%	67%	73%

Table 1: Confusion matrix HMMs + OWE

We have conducted two experiments based on two different training experiments: one using the hand labelled segmentation, the other one using the constrained HMM segmentation. The table 1 gives the confusion matrix in the case of segmentation. When trained with the hand labelled segmentation, we observed a recognition rate 5% less than the system trained with the HMM segmentation. It can be explained by the fact that the HMM segmentation does not introduce any bias between training and testing. This experiment shows a average of 78 % of correct recognition rate on the 6 stop consonants (/p/, /t/, /k/, /b/, /d/ and /g/) with a homogeneous repartition of the recognition rates over the six consonants.

4.4.2 With unconstrained HMM segmentation

In this case the HMM provides a unconstrained string of phonemes that determines the segmentation. The HMM phone accuracy (#phonemes - #insert. - #delet. - #substit.)/#phonemes is 70.6% [6]. From the recognized string of phonemes, we extract the six phonemes /p,t,k,b,d,g/. On these segments the OWE provides a new classification. A simple rule based on *a posteriori* probabilities of classes mixes the two answers: if the *a posteriori* probability of the OWE output winner is greater than a given threshold (typically 0.5) then choose the OWE answer else choose the HMM answer.

This simple algorithm does not yield any significant improvement except for phoneme /g/ whose recognition rate increases of 10%. It is not a surprising result because /g/ is known as a plosive which is highly influenced by the right vocalic context which is captured by OWE.

5. CONCLUSION

We have presented a hierarchical phoneme recognition system based on HMM and OWE models. The temporal segmentation is carried out by the HMM and the OWE performs the context-dependent classification. So far, the OWE model only re-classifies the six plosives. The best results have been obtained when the training is done with HMM segmentation that does not introduce any bias between testing and training conditions.

We have shown the great capabilities of OWE architecture on context-dependent classification which is a difficult task in speech recognition. For instance, the accuracy given by OWE alone is slightly smaller than the HMM accuracy. But the OWE accuracies for each phonemes are more homogeneous.

Our ongoing work is to generalize this recognition on the whole phoneme set and we plane to investigate various strategies in order to mix the answers coming from the two different parts of the system.

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 theirs Paragon Intel parallel computers.

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. 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

3. 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.

4. Lee, K.F., and Hon, H.W “Speaker-Independent Phone Recognition Using Hidden Marken Models”, IEEE Trans. ASSP, 37 (11), 1989

5. Mari, J.F, and Haton, J.P "Automatic Word Recognition Based On Second-Order Hidden Markov Models" ICSLP Yokohama, Japan. 1994

6. 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.

7. Mari, J.F., Fohr, D., Anglade,Y., and Junqua, J.C. "Hidden Markov Models and Selectively Trained Neural Networks for Connected Confusable Word Recognition". In ICSLP, pp S26-11. Yokohama, Japan. 1994

8. Pican, N., Fort, J.C, Alexandre, F. "A Lateral Contribution Learning Algorithm for multi MLP Architecture", in ESANN Proceedings. D Facto, Brussels, 20-22 April 1994.

9. Pican, N., Fort, J.C, Alexandre, F. "An On-line Learning Algorithm for the Orthogonal Weight Estimation of MLP". In Neural Processing Letters, D facto, Brussels, Vol 1, No 1, pp 21-24, 1994.

10. Pican, N., Alexandre, F.: "How OWE Architectures encode Contextual Effects in ANNs". In Mathematics and Computers in Simulation. 41 (5-6) July 1996.

11. Pican, N. :"Intrinsect and Parallel performances of the OWE Neural Network Architecture". In ICANN proceedings. Bochum, Germany. 17-19 July 1996.