Relatório Técnico

Núcleo de Computação Eletrônica

Automatic Speech Recognition: A Comparative Evaluation between Neural Networks and Hidden Markov Models

Antonio Carlos Gay Thomé
Suelaine dos Santos Diniz
Sidney Cerqueira Bispo dos Santos
Dirceu Gonzaga da Silva

NCE - 14/99

Universidade Federal do Rio de Janeiro

Automatic Speech Recognition: A Comparative Evaluation between Neural Networks and Hidden Markov Models

Antonio Carlos Gay Thomé – PhD
thome@nce.ufrj.br
Tel: 55 21 598 31 63
Dep. Ciência da Computação – IM-UFRJ
Ilha do Fundão, Cidade Universitária
20001-970 Rio de Janeiro – RJ – Br, 1998

Sidney Cerqueira Bispo dos Santos - DC
sidney@aquarius.ime.eb.br
Tel/Fax: (021) 542 09 99
Dep. Engenharia Elétrica – IME –DE/3
Pça Gen Tibúrcio, 80 - Praia Vermelha
22290 - 000 Rio de Janeiro – RJ – Br, 1998

Suelaine dos Santos Diniz – MC
suelaine@fund.cepel.br
Tel: 55 21 598 23 48
Dep. Ciência da Computação – IM-UFRJ
Ilha do Fundão, Cidade Universitária
20001-970 Rio de Janeiro – RJ – Br, 1998

Dirceu Gonzaga da Silva
dirceu@aquarius.ime.eb.br
Tel/Fax: (021) 542 09 99
Dep. Engenharia Elétrica – IME –DE/3
Pça Gen Tibúrcio, 80 - Praia Vermelha
22290 - 000 Rio de Janeiro – RJ – Br, 1998

Abstract - In this work we do a comparative evaluation between Artificial Neural Networks (RNA's) and Continuous Hidden Markov Models (CDHMM), in the framework of the recognition of isolated words, under the constrain of using a small number of features extracted from each voice signal.

In order to accomplish such comparison we used two models of neural networks: the Multilayer Perceptron (MLP) and a variant of the Radial Basis (RBF), and some HMM models. We evaluated the performance of all models using two different test sets and observed that the neural models presented the best results in both cases. Seeking to improve the HMM performance we developed a hybrid system, HMM/MLP, that improved the results previously obtained with all HMMs, and even those obtained with the neural networks for the all previous HMM, and even the neural nets for the hardest test set case.

1. Introduction

The project of any voice recognition system requires special attention with two major figures: the acoustic and the temporal components of the signal. It is known that the Markov models (HMMs) are able to tackle with both phenomena and that the same does not happen in the case of feedforward neural network types.

In spite of doing the acoustic modeling through the states mixtures and the temporal modeling through the transitions among states and the duration of the states, Markov models suffer from some peculiar problems. The independence hypothesis, for example, ignores any context information, and the first order Markov hypothesis may causes unsuitable modeling of the coarticulations and the duration of the states.

On the other hand, although they do not accomplish the temporal modeling, feedforward neural nets do present some other important characteristics such as: ability to learn complex functions, generalization, parallelism and tolerance to noise, that return them an important option for the area of voice recognition.

The main goal for this work was to compare the voice recognition performance among the two chosen neural models, the MLP and the Radial Basis nets, and some Continuous Density Hidden Markov Models (CDHMM). A reduced Portuguese vocabulary was chosen as used in [5], which is composed by the following words: liga (turn on), pare (stop), grave (record), pausa (pause), avance (move), siga (proceed), volte (return), desliga (turn off), ejete (eject) and apague (erase). The amount of 1697 repetitions of these words where recorded from a group of 113 different male speakers, averaging of 25 years old.

According to [5,9], we also limited the member of voice signal features to be used as input to the models, to maximum of 5. This constrain was adopted in order to reduce processing time and memory space requirements, mainly for the neural models.

2. The Models

2.1 HMM

We only used CDHMM - Continuous Density Hidden Markov Models for this work. The chosen structure for the models was the Baki's model, a particular case of the left-right, and for the training procedure, we adopted the Segmental-Kmeans method as described in [1].

For the experiments we chose 3 basic models as described below in Table 1.

Table 1 - HMM chosen for the comparative study

Choosen Models
8 states 5 gaussian mixtures (8s5m)
8 states 10 gaussian mixtures (8s10m)
15 states 10 gaussian mixtures (15s10m)

2.2 Neural Network

Two different models were used, the first one was a the traditional MLP with two hidden layers and learning rule implemented through the backpropagation algorithm [5]. The net network topology was composed as 481x69x20x10, i.e., 481 nodes in the input layer, 69 neurons in the first hidden layer, 20 in the second and 10 in the output layer. We adopted an adaptive learning rate, a momentum of 0.75 and the sigmoid function for all neurons. With this configuration the MLP network reached to an error of 2.7662 after 36710 epochs.

The other neural model used was a variant of the radial basis as developed in [5]. Its structure is basically composed by 4 layers divided into 2 stages, as illustrated in Figure 1.

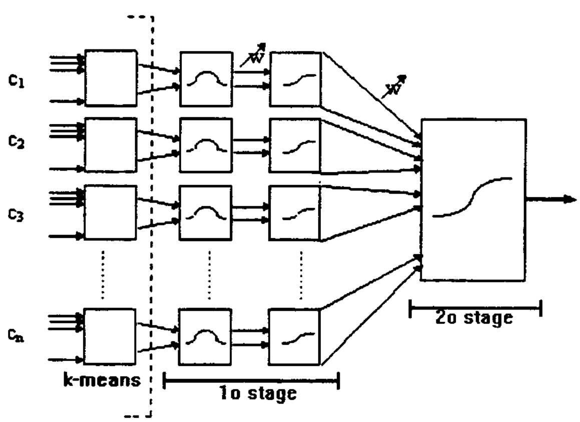

Figure 1 - Variant of the radial basis network

The first layer is only used for the training purpose, it is off line and performs an unsupervised clustering on the data set through the usage of the K-means algorithm. In our case, each pattern was composed by 4 sets of 120 points each corresponding each one, to one of the 4 voice features used as input and extracted over the 120 windows of the signal. Each of these data sets was divided into 10 clusters, giving each one the parameters, mean and deviation, to the construction of the second layer, that is composed by a set of gaussian functions. For the third and fourth layer of the model we used two 4 layer MLP networks, the first with a topology of 10x40x20x10 and the second of 40x55x20x10.

The training procedure is accomplished in three distinct phases. The first phase is off line and as said before, it consists of an unsupervised clustering over the data patterns. The second phase deals with the learning of the first set of MLP networks and the third phase ends the system learning through with the second MLP network. All MLPs were trained with the backpropagation algorithm.

For the test procedure it is only used the first stage compounded by the gaussian and first MLP layer, and the second stage (second MLP). In order to get a more detailed description of the model, see [5].

2.3 Hybrid System : HMM-MLP

In the CDHMMs, the durations of the events are modeled through the probabilities of the transitions, which gives to an exponential distribution.

Another way of getting information about the duration of the states is using the Viterbi's algorithm, to perform a direct measure of the training data segments as purposed in [1].

Using Viterbi's algorithm, the procedure for the recognition becomes the following: first the best segmentation is found and the corresponding likelihood is computed; after that the number of segments that were captured by each state is taken and the corresponding probability function is then estimated. The system likelihood, as can be seen in Eq 1, is then computed by the addition of the likelihood provided by the Viterbi's algorithm and the estimated probability function.

$$\log \hat{P}(q,O \mid \lambda) = \beta \log \mathrm{P}(\mathrm{q},\mathrm{O} \mid \lambda) + \alpha_d \sum_{j=1}^{N} \log[\mathrm{p}_j(d_j)] \quad (1)$$

where α and β are scaling multiplier, and d_j is the computed duration of the state j.

In order to estimate the probability density function as well as the value of the two scaling factors, and also to discriminate the input patterns among the HMMs, we projected the usage of a MLP network, as can be seen in the Figure 2. The network, receives as input from each of the 10 HMM, the duration of states and the likelihood computed by each HMM. For the hybrid system we used the "8s5m" HMM (table1) and a "90x20x10" MLP network.

For the network training we adopted an adaptive learning rate, starting it at the value of 0.002, a momentum factor of 0.7, and a sigmoid function for the neurons. The convergence occurred very fast, within 1500 epochs, to an error level of 0,001.

harder all models lost performance: around 16% the MLP, 18% the Radial Basis, 14% the HMM, and 11% the hybrid system, which proved to be the most robust model. This suggests that the combination of HMM and neural models may result in better systems, possibility because they may be able to take advantage of what is more effective in each its components.

TABLE 5: The tests results accomplished with RNAs, HMM and Hybrid System in % of recognition

Test	MLP	RBF	HMM	Hyb Syst.
Group I	95,0	**98,0**	87,3	93,6
Group II	79,0	80,5	73,5	**83,1**

4. Conclusions

In spite of Markov Hidden Models had not showed the same level of accuracy of the neural models, the experiments were valid to show that first ones are more robust and faster trained then the second.

The experiments also permitted us to conclude about the potentiality of hybrid models, especially those with the type of combination developed here, where the HMM takes care of the temporal modeling and the neural net takes cases of the acoustic modeling and context extraction. The state duration modeling provided by the neural nets was more efficient than that given by the exponential distribution of the standard HMM algorithm.

Another factor observed was the importance of the selection of a set of relevant features to represent the voice signal. It was possible to verify that the features used and chosen by [5] for the neural net systems, were not the most adequate ones for the case of the HMM systems. A larger set of features was not considered in our study because this would imply in a severe increase in of the complexity and time spent for training the neural models.

Our goal is to proceed the research evaluating different schemes of hybrid systems connecting HMM and neural nets, and also testing them over more complex data sets.

References

[1] RABINER, L. R. & JUANG, B.H. "Fundamentals of Speech Recognition" Prentice Hall, Inc., Englewood Clifs, Nova Jersey, 1993.

[2] DELLER Jr., J. R. et alii. "Discrete-Time Processing of Speech Signals", Macmillian Publishing Company, Nova Iorque, 1993.

[3] LEE, K. F. "Automatic Speech Recognitin — The Development of the SPHINX System", Kluwer Academic Publisher, Boston, 1989.

[4] PARANAGUÁ, E.D.S. "Reconhecimento de Locutores Utilizando Modelos de Markov Escondidos Contínuos", Tese de Mestrado, IME, 1997

[5] DINIZ, S.S. "Uso de Técnicas Neurais para o Reconhecimento de Comandos à Voz", Tese de Mestrado, IME, 1997

[6] HAYKIN, S. "Neural Networks - A Comprehensive Foundation", Macmillan 1994

[7] LEE,C.H; SOONG F.K.; PALIWAL K.K. "Automatic Speech and Speaker Recognition - Advanced Topics", Kluwer Academic Publisher, Boston, 1996.

[8] COLE, R. A. "Survey of the State of the Art in Human Language Technology", Center For Spoken Languagem Understanding, Oregon Graduate Institute, Publicações Técnicas, Nov. 1995

[9] DINIZ, S. S. , THOMÉ, A. C. G. "Uso de Técnica Neural para o Reconhecimento de Comandos à Voz", IV Simpósio Brasileiro de Redes Neurais, pp 23-26, dez 1997

[10] DUDA, O.R., HART, E., P., "Pattern Classification and Scene Analysis", Wiley-Intercience, 1973, pp. 114-118

[11] RENALS, S., MORGAN, N., "Connectionist Probabilit Estimation in HMM Speech Recogation", International Computer Science Institute, Berkeley, Dec 1992

[12] CHO, S-B, "Neural- Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals" , IEEE Transactions on Neural Networks, vol. 8, n.1, January 1997.