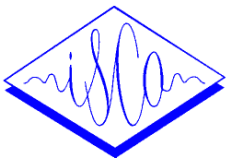
ISCA Archive
http://www.isca-speech.org/archive



4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

INCORPORATING FUZZY MODELLING IN A HYBRID HMM-ANNs SYSTEM FOR CSR TASKS

Xavier Menéndez-Pidal *', R. de Córdoba *, J. Ferreiros *, J.M. Pardo *
* GTH, Dpto IEL, ETSI de Telecomunicación, UPM, Spain
' ASEL, Univ. of Delaware - A.I. duPont Institute, USA
E-mail: pardo@die.upm.es or menendez@asel.udel.edu

Abstract

In this work, we present a new strategy to combine neural networks with HMMs which tend to take advantage of the modelling abilities of two independent modules, the ANN and the HMM, to make the design of a hybrid system less complicated. This approach incorporates fuzzy probabilistic information in the HMM to decompose the training task of a hybrid system. Using this strategy the training system is optimized about 7 times without significant loss of information. Also, we describe different techniques to improve the performances of the system which reduce the word error rate by 40%. Using this methodology, the hybrid system is trained much faster and can now benefit from two distinct sources of improvements such as neural modelling and classical HMM modelling which is less costly to perform.

10.21437/Eurospeech.1995-408

1. PROBLEM DESCRIPTION

One of the main problems with using ANNs as MLPs in CSR tasks is the time required to train the ANN on a large data base. To optimize the training procedure, many different ANN strategies have been proposed in the past years to improve the training time of a hybrid ANN-HMM system such as phoneme classification decomposition, extraction of acoustic information from independent ANNs, use of fast training ANNs criterium, and faster ANN training algorithms. However, training a complex ANN on a large data base is still highly computationally consuming. One of the most critical aspects of the classical hybrid implementation is the use of the ANN to perform all the modelling processes. While the HMM performs only a time alignment process computed in a very fast way, but it does not use the modelling capabilities of the classical HMM. The goal of this work is to determine the possibility of completing and dividing the training task in a hybrid system in both the ANN and the HMM modules. In our implementation, we minimize the training time "not training the ANN" and we complete the overall training system using the HMM as a modelling tool. This means that what is not done by the ANN is later accomplished in the HMM. This work takes advantage of a Dynamic Programming system which learns and codifies the knowledge by 2 or 3 orders of magnitude faster than the Back-propagation algorithms. This paper presents some experiments over 3 speakers (JWS04, BEF03 and CMR02) of the DARPA-RM1 data base where the training procedure of hybrid ANN-HMM system is performed by both modules incorporating fuzzy probabilistic modelling in the HMM in a similar way to the work performed in [6].

2. SYSTEM OVERVIEW

The basic main system is composed of an FFT preprocessor which produces 22-FFT spectral coefficients in the Mel-Scale every 10 ms, a modular hierarchical TSNN [8] neural network which recognizes 47 phones, an HMM-ANN training system (based on the Viterbi Algorithm) and an HMM-NN recognition system (based on the One-Stage algorithm [9]). For the initial ANN training process, we use automatic labelling performed by a previous DHMM. To drastically reduce the training procedure, the ANNs are trained with only one centered example for each phoneme in the input buffer of the ANN. In this way, the neural training could be accomplished in about 50 CPU-hours on a Sparc-Workstation training seven times less examples for each phoneme on average [2,3,8]. As the ANN performs a partial phoneme discrimination, the training system needs to be completed by including a kind of HMM modelling in the hybrid system [8]. This can be done in several ways. Recently many alternatives of inclusion of HMM modelling in the hybrid system have been proposed [2,3,4,6,8] to overcome the behavior of the ANN or improve the performances of a hybrid system. Here, we present a work where the inclusion of HMM modelling in a hybrid system has been performed to complete the training system. The most efficient and theoretical way to post-process and to model an ANN in an HMM seems to be when the ANN is interpreted as a VQ [6,7,8]. Using the ANN as a VQ, many simplifications could be carried out in a hybrid system without strongly increasing the complexity of that hybrid system.

3. THE FPI (Fuzzy Probabilistic Integration) HYBRID APPROACH

3.1. Extracting Centered Acoustic-Phonetic Indexes from an ANN-VQ

A modular ANN architecture has been chosen to parallel the ANN training as suggested in [2,3]. In this way, the ANN phoneme training is decomposed into smaller tasks which are easier to perform. In the hierarchical ANN system, 10 independent sub-nets (glides, nasals,..., broad-class) have been trained in a centered mode, covering 7 acoustic frames in each sub-net [3,8]. Training only the central part of the phoneme with the ANN, the behavior of the ANN is not known a priori in the transitional phoneme areas, but it is related to the centered phoneme features learned during the training. This means that the ANN can no longer be used as a global phoneme probability estimator, as proposed by Bourlard [1] to build-up a hybrid system [8]. Nevertheless, the sum of the outputs still tend to be close to one before any normalization, and the ANN could be interpreted as a Vector Quantizer of centered phoneme indexes. To reconstruct the centered acoustic-phoneme indexes from a hierarchical ANN trained in a centered mode, it could be performed as it was done for a global phoneme hierarchical ANN classifier in [2,3], (see Eq. 1).

$$P(fc/O) = P(fc/Cci, O) \cdot P(Cci/O) \quad (1)$$

Where:

P (Cci / O) is the a-posteriori probability of the acoustic-centered broad-class **Cci** given the observed O

P (fc / Cci, O) is the a-posteriori probability of the acoustic-centered phoneme **fc** in the sub-network **Cci** given the observed O

As the DARPA-RM1 data base is phonetically not balanced, the priors have been used to estimate the probability of each acoustic-centered phoneme, **P(O/fc)**, as suggested in [1]. These probabilities are obtained from the Bayes rule (see Eq. 2), dividing the a-posteriori probabilities by the priors and performing a final normalization to eliminate the term, **P(O)**, which is centered phoneme indexclass independent.

$$\frac{P(O/fc)}{P(O)} = \frac{P(fc/O)}{P(fc)} \quad (2)$$

3.2. Building the Probability of emitting an Observation in an HMM State

In the hybrid method proposed, Fuzzy Probabilistic Integration (FPI), the outputs of the ANN are not physically linked to any HMM model as the ANN outputs cannot be associated directly with a phoneme, but rather the HMM uses a fuzzy-confusion vector for each HMM state to determine the amount of information carried out by the N-best ANN's outputs. The confusion matrix components determine the average activation likelihood **Pm(fc/Jh)** of the ANN output **fc** for the **h** HMM state of the phoneme **J**, (see Eq. 3).

$$Pm(fc/Jh) = \frac{\sum_{\forall O \in Jh} P(O/fc)}{\sum_{\forall O \in Jh} 1} \quad (3)$$

Combining the N-best ANN outputs, **P(O/fc)**, with their respective value produced by the confusion matrix **Pm(fc/Jh)**, an approximation of the conditional probability **P(O/Jh)** is computed to obtain an observation **O** in an HMM state **Jh** [5] (see Eq. 4).

$$P(O/Jh) \approx \sum_{fc}^{5best} P(O/fc)\, Pm(fc/Jh) \quad (4)$$

The number of ANN outputs to be used in the HMM is task dependent. The information brought by the N-best outputs are dependent on the overlap of the phonemes to discriminate. For the DARPA-RM1 task, the acoustic information is distributed along the five best outputs. Using the five best outputs, the probability of being in an HMM state could be estimated as in Eq. 4 without loss of information. If we use less number of outputs, the accuracy of the system degrades, as the hybrid system does not extract the global information brought by the VQ. Using more outputs does not achieve improvement, while we also found it increased the overall recognition computational cost. Performing this multi-labelling process in the HMM, the combination of the confusion matrix and the 5-best ANN outputs has allowed us to estimate an approximation of generating an acoustic observation **O** in the state **h** of the phoneme **J** without performing an explicit discrimination with the ANN of all the **h** state.

3.3. Initialization of the Viterbi Algorithm and HMM Topology

3.3.1. Initial acoustic values of the HMM models

The initial value of the confusion vectors for each **h** state of the phoneme **J** has been estimated from the theoretical behavior of the ANN expected for a lexical position as is shown in Eq. 5.

$$Pm(fc/Jh) = \begin{cases} 1, fc = Jh, \forall h \\ 0, fc \neq Jh, \forall h \end{cases} \quad (5)$$

Using these initial values, the first alignment performed in the FPI approach is the same as in the Bourlard and Wellekens scheme[1], and iteratively the confusion vectors are estimated using Eq. 3. In our experiments the systems requires 3 iterations of the Viterbi recursion to perform the final training which is accomplished in 2 hours in a Sparc-Workstation.

3.3.2. HMM topology

As the ANN inserted centered phonetic symbols unknown in the transitional zones of the phonemes, a long HMM topology has been used to avoid the production of inserted phonemes in the FPI recognition system. Using 5 states per phoneme in the HMM, a trade off between inserted and deleted phonemes is obtained, minimizing the effect of those transitional symbols emitted by the ANN. The centered phonetic HMM topology shown in Fig. 1 provides good accurary, avoiding phoneme insertion and modelling correctly the ANN symbols produced in the phonetic transitions.

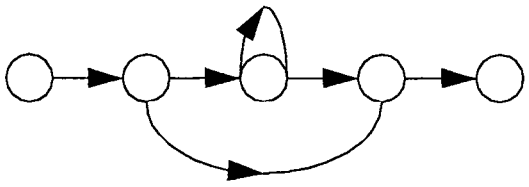


Fig. 1, 5-state HMM phoneme topology

4. RESULTS

Table 1, illustrates the results obtained from the three speakers with and without using the "official" word-pair in the One-Stage recognition system, with exactly the same system implementation.

Table 1: Word Error, Perplexity 991 and 66, for the three speakers with the initial system

Speaker	WE, P-991	WE, P-66
JWS04	26.4%	4.5%
CMR02	32.5%	6.9%
BEF03	42.1%	11.9%
Average	**33.7%**	**7.8%**

Different techniques have been accomplished over the two speakers (CMR02 & BEF03) with worse scores to improve their accuracy. Experiments to correct the initial labelling provided by the DHMM have reduced the system errors. Retraining the ANN over the new alignments provided by our FPI hybrid system have improved at frame level and at word level the performances of the system as proposed by Bourlard [1]. The greatest improvements have been observed above all in the first iteration (see Table 2). To improve the system, multiple sources of information have also been introduced in the hybrid system in an independent way as suggested by Le Cerf [6,7]. We have introduced multiple independent hierarchical ANN-VQs trained on different sets of parameters in a centered mode (22-FFT coefficients, 8 delta-FFT coefficients, 8 delta2-FFT coefficients) to further decompose the ANN training task in small independent nets. The following table illustrates the improvements obtained with these different techniques in our hybrid One-Stage HMM-ANN system.

Table 2: Evolution of the Word Error for speakers CMR02 and BEF03, after realignment and incorporating Multiple independent Hierarchical ANN-VQ trained on different sets of Features (fft, d-fft & d2-fft), for Perplexity 991 & 66

System	CMR02	BEF03
1st, align.	26.9%-5.6%	38.2%-9.3%
2nd, align.	26.2%-5.4%	37.1%-8.9%
2 ANN-VQ	23.1%-4.6%	35.4%-7.5%
3 ANN-VQ	21.6%-4.2%	34.6%-6.8%

Also, the evolution of the complexity of the FPI system

when using more ANN-VQ is depicted in Table 3.

Table 3: Number of free parameters per speaker, in the ANN and in the HMM, for the FPI hybrid systems for the different spectral sets of information added

System	1 ann-vq	2 ann-vq	3 ann-vq
#ann+hmm	28+12 K	44+24 K	54+36 K

The last configuration system is an example of highly decomposed phoneme discrimination task. Instead of using 1 global network to discriminate all 47 centered phoneme indexes, 30 independent nets have been used per speaker, taking advantage of the modular ANN principle trained on independent feature sets. Also, this system take advantage of the classical modelling abilities of the HMM to reconstruct the probability **P(O/Jh)** of emitting an observation **O** in an **Jh** HMM state using independent Fuzzy ANN-VQs, computing **P(O/Jh)** as follows (see [5] & Eq. 6).

$$P\left(\frac{O}{Jh}\right) \approx P\left(\frac{O}{Jh}\right)_{(fft)} P\left(\frac{O}{Jh}\right)_{(d-fft)} P\left(\frac{O}{Jh}\right)_{(d2-fft)} \quad (6)$$

5. CONCLUSIONS

Using this new methodology, the hybrid system could be strongly optimized in training time, but it also could be improved. In this paper, we have illustrated two different techniques to improve the system correcting the initial labelling with the hybrid system and incorporating more acoustic information in the FPI scheme. Comparing the global results obtained here with those obtained at Limsi-Phillips or at ICSI [1,2,3] in the DARPA-RM1 data base, the centered ANN training has not degraded the accuracy of the system. The FPI scheme tends to give the same results as the Bourlard scheme, performing a trade off between training and recognition requirements. In an acoustical point of view, even if the phoneme transitions are not trained with the ANN, the behavior of the ANN in the transitional zones seems to be consistent and the ANN has learned enough information to build a speech recognition system without a significant loss of accuracy. In this hybrid strategy based on a cooperative perspective between the ANN and the HMM system in the modelling task, the training procedure could be divided and performed in both modules in the hybrid system. But this strategy could also be used to improve the accuracy of a final hybrid system, taking advantage of the modelling abilities of classical HMM which are faster and easier to perform. For example, the HMM modelling could be improved performing a context-dependent modelling as the ANNs are now used as classical Fuzzy-VQs [5] post-process in the HMM.

Acknowledgments

This work has been realized in the GTH "Grupo de Tecnologia del Habla", supported by the Spanish CICYT grant number TIC94-0119 and partially by the Spanish Ministry of Education and Science grant number FPI-PF94-51.368.925.

REFERENCES

[1]**Bourlard, H., Morgan, N.** *Connectionist Speech Recognition a Hybrid Approach*, Kluwer Academic Publishers, 1994.

[2]**Devillers, L.** "Reconnaissance de parole continue avec un système hybride neuronal et markovien", *Thèse Université de Paris-Sud, Centre D'Orsay*, Novembre 1992.

[3]**Dugast, C., Devillers, L., Aubert, X.,** "Combining TDNN and HMM in a Hybrid System for improved continuous speech recognition", *IEEE Transactions on Speech and Audio Processing*, 1994, vol II, No. 1, pp 217-223.

[4]**Driancourt, X.D., Gallinari P.** "A speech recognizer optimally combining learning vector quantization, dynamic programming and multi-layer perceptron", Proc. of *ICASSP-92*, vol I, pp 609-612, 1992.

[5]**Huang, X.D., Ariki, Y., Jack, M.A.** *Hidden Markov Models for Speech Recognition*, Edinburg University Press, 1990.

[6]**Le Cerf, P., Van Compernolle, D.,** "MLPs as Labelers For HMMs", *IEEE Transactions on Speech and Audio Processing*, 1994, vol II, No. 1, pp 185-193.

[7]**Le Cerf, P., Van Compernolle, D.,** "Using parallel MLPs as Labelers For multiple codebook HMMs", *ICASSP-93*, pp 1561-1564.

[8]**Menéndez-Pidal, X. Ferreiros, J. Córdoba, R., Pardo, J.M.** "Recent Work in Hybrid Neural Networks and HMM systems in CSR tasks", Proc. of *ICSLP-94*, vol 3, pp 1515-1518, 1994.

[8,b]**Menéndez-Pidal, X.** "Contribución al Reconocimiento del Habla con Redes Neuronales y Algoritmos de Programación Dinámica", *Tesis Doctoral, ETSIT-UPM*, Marzo-1995.

[9]**Ney, H.** "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol 32, pp 559-562, 1984.