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

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A HYBRID ANN/HMM SYLLABLE RECOGNITION MODULE BASED ON VOWEL SPOTTING

John Sirigos, Nikos Fakotakis and George Kokkinakis
Wire Communications Laboratory
University of Patras, 26500 Greece
john@pat.forthnet.gr

ABSTRACT

This paper presents a hybrid ANN/HMM syllable recognition module based on vowel spotting. An advanced multi-level vowel spotting method is used to achieve minimum vowel loss and accurate detection of the vowel location and duration. Discrete Hidden Markov Models (DSHMM), Multi Layer Perceptrons (MLP) and Heuristics (HR) are used for this purpose.

A hybrid ANN/HMM technique is then used to recognize the syllables between the detected vowels. We replace the usual DSHMM probability parameters with combined neural network outputs. For this purpose both context dependent (CD) and context independent (CI) neural networks are used. Global normalization is employed on the parameters as opposed to the local normalization used on parameters in standard HMMs. Also, all parameters are estimated simultaneously according to the discriminative conditional maximum likelihood (CML) criterion. The tests were performed on the TIMIT and NTIMIT databases and showed significant performance improvement compared to similar systems.

1. INTRODUCTION

While it is true that ASR is a hard task and that ANNs can be helpful for hard pattern recognition problems, we are wary of assuming that neuralware can implement a complete ASR system. Practical pattern recognition tasks are rarely implemented by a monolithic element, either in the form of a single ANN or any other homogeneous component. In particular, we have not yet found a way to use neural networks to implement a complete system for the recognition of continuous speech. However we have learned to use an ANN as a key component in such a system [1].

In order to improve the neural network estimation we can use an ensemble of neural networks [2], hereafter called a neural combination. In theory it is proved that the combination generalization error; is less than the (weighted) average of member-network errors the smaller the correlation between the member networks, the smaller the combination error.

Regarding hybrid systems, previous work by Bourlard and Morgan [3] showed that both in theory and actual practice multilayer perceptrons (MLP) can be successfully used in Hidden Markov Model (HMM)-based speech recognition for estimating the state-dependent observation probabilities. There are several ways of making ANNs and HMMs cooperate in hybrid systems. Numerous of studies have concentrated on the use of ANNs as front-ends for HMMs. It has been shown that a properly trained MLP for pattern classification is asymptotically equivalent to an estimator of a posteriori class probability.

Hybrid systems like the one mentioned above have been reported to enhance the recognition performance of HMM systems. Another approach to hybrid system design considers ANNs as HMM postprocessors. This allows the efficient combination of the time-alignment capability of HMMs with the discriminative power of ANNs, a feature, which is of particular relevance to continuous speech recognition. Some systems feed sentences recognized by the HMM into the ANN to perform a second stage discrimination. This solution works only for applications requiring a limited vocabulary.

In this paper we present a syllable recognition system based on vowel spotting and on the use of a hybrid ANN/HMM with combined neural networks. We denote as a syllable the letters between two vowels along with the second vowel.

Our approach for modeling each syllable is similar to the idea of adaptive input transformations, but we propose to replace some or all HMM probability parameters with the output of small state specific combined neural networks. Simultaneous estimation of all parameters is then performed, similar to what is done for adaptive input transformations. A proper probabilistic interpretation is guaranteed by normalizing the model globally as opposed to the often-approximate local normalization enforced in many existing hybrids. We avoid the calculation of the normalization term by using the discriminative CML criterion for training.

The final output of the whole system is the sequence of the recognized syllables of the spoken sentence. The system was trained and tested with the TIMIT and NTIMIT databases. It is important to say that no grammar or lexicon was used in the overall system.

2. SYSTEM DESCRIPTION

The block diagram of the overall system is presented in figure 1. In the first stage pre-processing of the speech signal, takes place using the RASTA-PLP method. In the second stage, vowel spotting is performed by measuring the location and the duration of the vowels on the pre-processed speech signal, using a multi-level technique [4]. For this purpose neural networks, discrete HMMs and heuristics are used, controlled by an appropriate unit. The main objective of this multi-level technique is to minimize the possibility of a vowel loss.

The third stage is a hybrid MLP/HMM system that is used to recognize the syllables between the detected vowels and also try to correct possible vowel spotting errors. Combined neural networks consisting of multi layer perceptrons along with

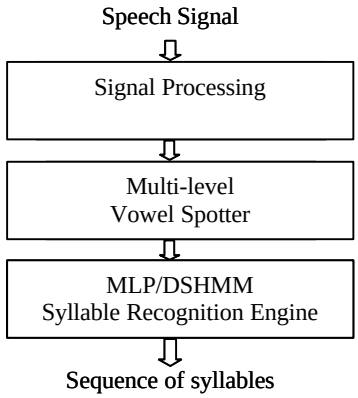

Figure 1. Overall system diagram

discrete Hidden Markov Models are used.

2.1 Signal Processing

Speech input is sampled at 16 kHz, digitized at 16-bit resolution and segmented into frames of 22 ms duration with 9 ms overlapping. After applying a Hamming window each frame is analyzed using the relative spectral perceptual linear predictive (RASTA-PLP) parameterization to obtain the characteristic parameters of the signal. Then cepstral recursion follows to compute the cepstral coefficients of the RASTA-PLP model.

Thus, at every q^{th} frame a parameter vector of 12 cepstral coefficients, the corresponding 12 delta coefficients and 1 log energy coefficient are calculated to form the 25 feature vector.

2.2 Vowel Spotting

For detecting the vowels, we use a multi-level combination of three different classification stages controlled by a unit which selects each time the proper method and directs the flow of the vowel spotter: Multi Layer Perceptrons (Stage 1 - MLP), Hidden Markov Models (Stage 2 - HMM) and Heuristics Rules (Stage 3 - HRULE). Each stage is linked with a set of rules [4] that are used by the control unit in order to select the appropriate stage. Figure 2 shows the block diagram of the vowel spotter.

The MLP employed is a 3-layer (2 hidden layers) feedforward artificial neural network. The classification output of the network is passed through a three-point median filter to eliminate isolated impulse noise. The HMM, used in the second-stage, consists of three different models for each vowel: one for the middle part of the vowel, one for the left and one for the right. The third stage consists of a set of heuristics rules (HRULES) based on previous work [5]. Vowel candidates' location is performed on the smooth energy function by using a peak-picking procedure. Then ripple rejection and strong consonant rejection follows.

The control unit takes as input a sentence from the pre-processing unit and activates the 1st stage. The output of the MLP, after passing through the median filter, is fed back to the control unit. Then by using the set of rules linked with this stage, the control decides whether there is a falsely rejected vowel or a falsely accepted non-vowel phoneme. When a mismatch of this kind is detected, it proceeds to the second stage, otherwise it proceeds with the following input sentence. When the second stage is activated, the control unit passes the mismatch part of this sentence to the HMM stage. Both MLP and HMM outputs along with the set of rules for the HMM, are used to calculate a "success" factor. If this factor is greater than 1 then the control continues with the next sentence, otherwise it goes to stage 3. When the third stage is activated, which takes as input the mismatch region (calculated by the two previous stages), all the different outputs of the three stages along with a set of rules connected to the heuristics are used to calculate a new "success factor". If it exceeds 1, the control proceeds with the next sentence otherwise it raises fatal error, which means that there is a doubtful region in the sentence.

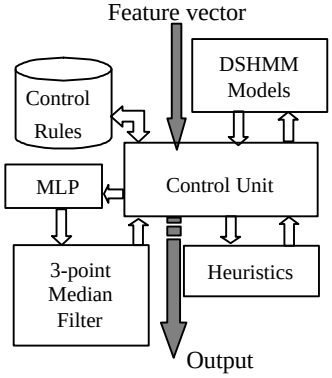

Figure 2. Vowel Spotter block diagram

2.3 Syllable Recognition Engine

The ANN/HMM syllable recognition engine takes as input the position and the duration of a vowel along with the distance to the next located vowel. For this time period in the speech signal, the recognizer is called to find the corresponding syllable.

The procedure is as follows: The feature vector is passed through the vector quantization module (VQ) which translates the 25 features into two integer numbers and the log energy coefficient. We use one VQ for the 12 cepstral and one for the 12 Ä-cepstral coefficients. So a 3 number vector is fed to the combined network. The estimated probabilities by the network are fed to the discrete HMM models to detect the proper syllable.

Afterwards the threshold-distance checking takes place. The outputs of the vowel spotter and of the hybrid model are taken into account in order to correct possible errors of the vowel detection. Finally a syllable appears in the output of the recognizer. Below, we describe each stage analytically.

2.3.1 Combined network description

In the proposed system the combination of a Context Independent (CI) and a Context Dependent (CD) ANN forms a combined network (CBANN). Since the CI network is usually more robust (due to more training data for each output), while the CD network is more precise (because of the detailed modeling of each context of the database phones), a combination may be both robust and precise. Theoretically it is proven that the use of combined networks improves the performance of a system [6].

Let X_{ci} be the output score for phone X of a CI network and Y_{cd} the output score for class Y of a CD network. Their combined output will be XY_{com}:

$$XY_{com}=aX_{ci}+(1-a)\sum_{Y\in Y} Y_{cd} \qquad (2)$$

2.3.2 Hybrid module description

We use the combined network along with discrete HMMs with the intention to create an engine for the recognition of the syllables between vowels. We use both match and transition networks for estimating the probabilities of the HMM models.

The HMM emission probability $\dot{O}_i(x_{in})$ of the observation vector x_{in} in state i is replaced by a match combined network $\ddot{O}_i(s_i;CBANN^1_i)$, which is parameterized by the weights of the combined network used for state i with input s_i and only one output. The neural network input s_i corresponding to x_{in} is a window of context around $x_{I,}$, a symmetrical context window of 2k+1 observation vectors, $x_{in-k},x_{in-k+1},\ldots, x_{in+k}$. Similarly, the probability of a transition from state i to j $\dot{E}_{ij}$, is replaced with the output of a transition network $\dot{E}_{ij}$ $(s_i,CBANN^2_I)$ which is also parameterized by another combined network. The transition network assigned to state i has J_i outputs, where J_i is the number of (non-zero) transitions from state i. In our system the neural networks being used are standard feed-forward MLPs but we can also use recurrent or radial basis networks.

In complete analogy with the likelihood p(x|M) of a HMM for observation sequence $x=x_1,\ldots x_L$, we define the quantity

$$q(x\,|\,M)=\sum_{\pi} q(x,\pi\,|\,M) \qquad (3)$$

with

$$q(x,\pi\,|\,M)=\prod_{i=1}^{L}\Theta_{\pi_{i-1}\pi_i}(s_{i-1};CBANN^2_{\pi_{i-1}})\Phi_{\pi_i}(s_i;CBANN^1_{\pi_i})$$

(4)

where M denotes the whole model and state sequence ð=ð1,…,ðL is a particular path through the model. We also define ð0=0 and È ii(s0;CBANN20) as the probability of initiating a path in state i.

So is the context we choose to associate with the beginning of the sequence. The probabilistic interpretation is ensured by explicit normalization of q,

$$p(x \mid M) = \frac{q(x \mid M)}{\int_{x' \in X} q(x' \mid M) dx'} \quad (5)$$

Each observation vector xi has an associated label yi corresponding to the class to which it belongs. Each class is a syllable. Similarly, each state is assigned to a class label. To maximize the prediction accuracy we choose parameters so as to maximize the conditional likelihood of the observed labeling y=y1,…,yL,

$$P(y \mid x, M) = \frac{p(x, y \mid M)}{p(x \mid M)} \quad (5)$$

Maximizing the above probability is known as Conditional Maximum Likelihood estimation (CML) and is equivalent to Maximum Mutual Information estimation (MMI) if the language model has been fixed during training. P(x,y|M) is calculated as a sum over all paths consistent with the labeling, i.e., if observation l is labeled f only paths in which the l^{th} state has label f are allowed. If the set of these consistent paths is called A(y) we have [7],

$$p(x, y \mid M) = \frac{q(x, y \mid M)}{\sum_{y'} \int_{x' \in X} q(x', y' \mid M) dx'} \quad (6)$$

with

$$q(x, y \mid M) = \sum_{x \in A(y)} q(x, \pi \mid M) \quad (7)$$

Since

$$\sum_{y'} q(x', y' \mid M) = q(x' \mid M) \quad (8)$$

the normalization is the same as in (4) and (6) so,

$$P(y \mid x, M) = \frac{q(x, y \mid M)}{q(x \mid M)} \quad (9)$$

and the normalizing factor has conveniently disappeared. Both q(x|M) and q(x,y|M) can be calculated by a straight-forward extension of the forward algorithm [7].

To maximize (9) we use stochastic online gradient ascent augmented by a momentum term, where the parameter update is performed after each observation sequence.

2.3.3 Threshold-distance checking

In case the vowel spotter raises an error that a vowel has been missed, then the hybrid is used to detect two syllables in the period the error occurred, by steps of 20ms. By using the top 3 recognized syllables by the hybrid and the vowel spotter's score along with the hybrid score we decide which of the 3 recognized syllables is the most correct.

If the error denotes that a vowel is added, then if the outputs of the hybrid fall between a pre-calculated threshold, the period is increased up to the next found vowel and the hybrid runs again to search for a syllable in this period. The threshold is given by the following equation: $Th_j = (1 - a_j) \cdot H_{\min j}$

where $a_j = \frac{1}{N} \sum_{i=1}^{N} \left(\left(\frac{1}{N} \sum_{k=1}^{N} H_{ij} \right)^2 - H_{ij}{}^2 \right)^{1/2}$ (10)

So both missed and inserted vowels are checked by this module in order for the performance to be increased. If after the described procedure is terminated there is still a missed or inserted vowel then the recognized syllable may be wrong. We still work on this problem to improve the technique so as to eliminate the missing and inserted vowels, which misguide the system.

3. EXPERIMENTAL RESULTS

3.1 Databases

For our experiments we used the TIMIT and NTIMIT databases. We chose them for our experiments because they are phonetically balanced, and in addition there are time-aligned phonetic transcriptions of all the sentences in the databases. Both databases contain 630 speakers (438 male and 192 female) each of them having uttered 10 sentences. Each utterance is a sentence of approximately 3 seconds duration.

3.2 Training

As a training set we used the 462 speakers of the TIMIT/NTIMIT training set. We used all the sentences in order to train our models for the syllables found. The tests were carried out by using the recommended TIMIT/NTIMIT core test set.

For training the vowel spotter we used a standard 4-state left-to-right model for the HMM states and a codebook size of 512 vectors. The vowel phoneme categories found on TIMIT/NTIMIT database were 20. For the MLP stage of the vowel spotter we used a 3-layer (2 hidden layers) feedforward artificial neural network. The overall architecture of the network, i.e., the number of hidden layers and the number of nodes per hidden layer, was determined experimentally by training the network. The experiments resulted to an optimal network size of 25x12x8x1. A fast version of the back propagation algorithm was used [8]. One sentence from each speaker, thus 462 sentences in total, was used for training the vowel spotter. The transcription files of the TIMIT database were used to calculate the duration and location of the vowels.

For training the DSHMM models of the recognition engine, we used a simple left-to-right five-state model. The last state in any submodel is fully connected to the first state of all other submodels. The choice of the number of states in the HMM is made empirically. The use of four states in our experiments gives satisfactory performance, which is either comparable or superior to the use of other state numbers.

Our baseline system is a standard discrete HMM using a codebook of 512 prototype vectors. Transition and state probabilities of the HMM are replaced, as described before, by combined networks with a symmetric input window of 2K+1 observation vectors, where K was set to 2 and a sigmoid output function was used. The Baum-Welch reestimation algorithm was used for ML training. The time required for training was less than 40 hours on a fast workstation for all models employed in the syllable recognition problem.

The size of the two ANNs, the CD and the CI network, which form the CBANN, was experimentally chosen to 25x25x1 (25 neurons for the hidden node) for the CD and 25x22x1 (22 neurons for the hidden node) for the CI.

3.3 Testing

We tested our system with the TIMIT database for clean speech conditions and NTIMIT for noisy conditions. For calculating the recognition accuracy the following equation was used:

%Accuracy=100%-%Insertions-%Deletions-%Substitutions

Syllable recognition experiments were conducted for various data sets corresponding to different dialect regions. The accuracy is shown in table 1 for both TIMIT and NTIMIT, with and without threshold-distance checking. In interpreting the classification results in this table, we can easily see the performance boost by using this module for both databases, which justifies its importance. Especially with NTIMIT the accuracy improvement is much more significant.

Data set	*Accuracy with threshold checking*		*Accuracy without threshold checking*	
	TIMIT	*NTIMIT*	*TIMIT*	*NTIMIT*
1	75.81	59.02	70.11	50.42
2	76.24	60.32	75.28	51.06
3	73.55	57.13	69.73	45.12
4	73.74	58.09	70.55	49.86
5	77.23	60.11	76.01	52.61
6	72.45	60.04	68.34	51.78
7	76.83	57.89	72.43	50.22
8	74.88	61.76	74.05	53.07

Table 1. Accuracy of the system with and without the threshold-checking module, for TIMIT and NTIMIT databases for various data sets.

Figure 3 shows the classification rate by changing the percentage of training data used for each data set of the TIMIT database. We can see that the performance of the recognizer is almost stable for different dialect regions of the same database.

It is of great significance to take a closer look on the error rate. We can divide this error rate into the falsely recognized vowels and the falsely recognized syllables. Although the threshold-checking module tries to minimize the falsely recognized syllables by using both the vowel spotter output and syllable probabilities, errors still occur. By using the proper grammar the performance of the system regarding the syllable error rate will be increased. The total error rate at the output of our system as a function of the training data contribution of the vowel spotter and the syllable recognizer is shown in figure 4.

4. CONCLUSIONS

In this paper we presented a syllable recognition system based on a hybrid ANN/HMM recognizer and vowel spotting. In our tests no grammar was used, as our objective was to maximize the performance of the signal processing part of a general-purpose speech recognition system. It has been shown that vowel spotting is a good way to segment the speech signal with good precision. Instead of using phone segmentation we use the syllables located between successive vowels. Also, for each syllable we combine the results of the vowel spotter and the syllable recognizer to improve considerably the performance of the whole system.

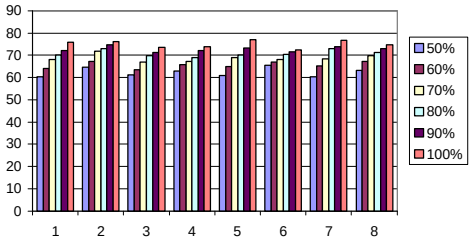

Figure 3. Recognitino accuracy for TIMIT database by changing the amount of training data.

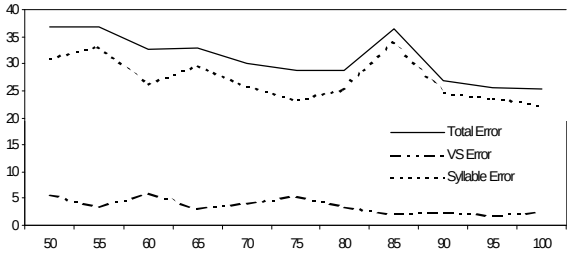

Figure 4. Accuracy of the system as a function of the training data for TIMIT database.

The following three elements contribute in making the system robust and precise:

a) Preliminary segmentation by using a high accuracy multi-level vowel spotter.
b) Combined neural networks along with discrete HMM for improved syllable recognition accuracy.
c) Combination of vowel and syllable probabilities to increase the overall performance ("Threshold-distance checking" module).

Our current plan is to test the system in other noisy environments by using the noise reduction technique described in [9] and show that the system can be also used in noisy environments without retraining.

5. REFERENCES

[1] Ha-Jin Yu, Yung-Hwan Oh, *A neural Network using Accoustic Sub-word units for Continuous Speech Recognition* , ICSLP 96, pp. 506-509, October 1996.

[2] L. Breiman, *Stacked Regressions* , Technical report, University of California, Berkeley, 1994.

[3] H. A. Bourlard and N. Morgan, *Connectionist speech recognition* , Kluwer Academic, Boston, MA, 1994.

[4] J. Sirigos, N. Fakotakis, G. Kokkinakis, *A high-performance vowel spotting system based on a multi-stage architecture* , Eusipco 98, Rhodes, Greece.

[5] N. Fakotakis, E. Tsopanoglou and G. Kokkinakis, *A Text-Independent Speaker Recognition System Based on Vowel Spotting* , Speech Communication Journal, Vol. 12, No. 1, pp. 57-68, March 1993.

[6] Brian Mak, "*Combining ANNs to improve phone recognition*", ICASSP 97.

[7] Soren Kamaric Riis, Anders Krogh, *Hidden Neural Networks: A Framework for HMM/NN Hybrids*", ICCASP 1997.

[8] T.P.Vogl, J.K.Mangis, A.K.Rigler, W.T.Zink and D.L.Alkon. *Accelerating the Convergence of the Back-Propagation Method* , Biological Cybernetics 59, pp. 257-263.

[9] J. Sirigos, N. Fakotakis, G. Kokkinakis, *Improving Environmental Robustness of Speech Recognition using ANNs* , DSP '97, Santorini, Greece.