Neural Network Preprocessor for Recognition of Syllables

A. Nayeemulla Khan, Suryakanth V. Gangashetty, and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, India.
email: {nayeem,svg,yegna}@cs.iitm.ernet.in

Abstract

The recognition rate of syllables in continuous speech is hampered due to the large size of the syllable vocabulary and the confusability among them. One approach to reduce the confusability and the search space is to preclassify the syllables into a small set of equivalent classes and then perform recognition within a particular equivalent class. In this study, the syllables in a language are grouped into equivalent classes based on their consonant and vowel structure. The syllables that map onto an equivalent class are called 'cohorts'. Artificial neural network models are used to preclassify the syllables into the equivalent class to which they belong. This is followed by recognition of the syllables among the smaller number of cohorts within a class by means of hidden Markov models. The preprocessing stage limits the confusable set to the cohorts within a class and reduces the search space. This hybrid approach helps improve the recognition rate over that of a plain HMM based recogniser.

1. INTRODUCTION

Hidden Markov models are capable of modeling the variability in the speech sound, and absorbing the variability in the length of the speech signal. Discriminative training is not provided in these models. Neural networks on the other hand provide discriminative training, and are good classifiers. They are capable of capturing the nonlinear discriminative surfaces separating the different classes. But they require fixed dimension patterns representing the speech sound. It is possible to combine the strengths of either models and develop hybrid models for enhanced performance.

In the task of developing automatic speech recognition systems for Indian languages, it is observed that, syllables seem to be the more appropriate unit for recognition [1][2]. The syllables in any language are in the order of a few thousands. They are dynamic sounds, and are highly confusable among each other. The performance of hidden Markov models (HMM) based recogniser degrades as the size of the subword vocabulary increases. Developing a monolithic artificial neural network (ANN) model to discriminate between all the syllables is infeasible due to the size of the resulting classifier and the varying frequency of occurrence of different syllables in continuous speech. Alternatively modular neural networks have been proposed to categorise a subset of the frequently occurring syllables in the language [2]. Constraints based on the confusability of the consonant-vowel units have been used to enhance the evidence in a constraint satisfaction neural network for stop consonant-vowels [3]. ANN models have also been used to classify words or phonemes by mapping the temporal representation of the sound into spatial representations [4]. In hybrid ANN-HMM networks, the ANN has primarily been used to estimate the posterior probabilities of the states of the HMM for improved performance [5].

In this paper we explore a method of enhancing the recognition rate of syllables extracted from continuous speech by preclassifying them using a neural network as a preprocessor. This preclassification step is followed by recognition of the syllables using conventional HMM models. The syllables are initially grouped into a set of equivalent classes based on their syllable structure. The syllables within an equivalent class are called 'cohorts'. A multilayer feed forward neural network (MLFFNN) model is trained to classify the syllable into the equivalent class to which it belongs. In the next stage a HMM based syllable recogniser is used to recognise the syllables within the equivalent class. The search space of the recogniser now reduces from the list of syllables in the vocabulary to the syllables within a particular equivalent class, thereby reducing the confusability and increasing the recognition rate. This two stage classification helps improve the performance of the recogniser. The performance of the MLFFNN based hybrid system is compared with an equivalent support vector machine (SVM) based hybrid system.

This paper is organised as follows. The next section gives the system description. Section 3 describes the experiments and the results of the study. The final section summaries the study.

2. SYSTEM DESCRIPTION

The block diagram of the proposed system is shown in Figure 1. For every syllable segment in continuous speech obtained by manually segmenting the signal, 13-dimension mel-frequency cepstral coefficients (MFCC) are extracted for every frame of 15 msec with a frame shift of 5 msec [6]. A fixed 130-dimension pattern vector, is formed by the concatenation of 10 frames corresponding to the syllable segment. These frames are chosen by the method of linear compaction and elongation as described in [7]. These large dimension

0-7803-8243-9/04/$17.00 © 2004 IEEE

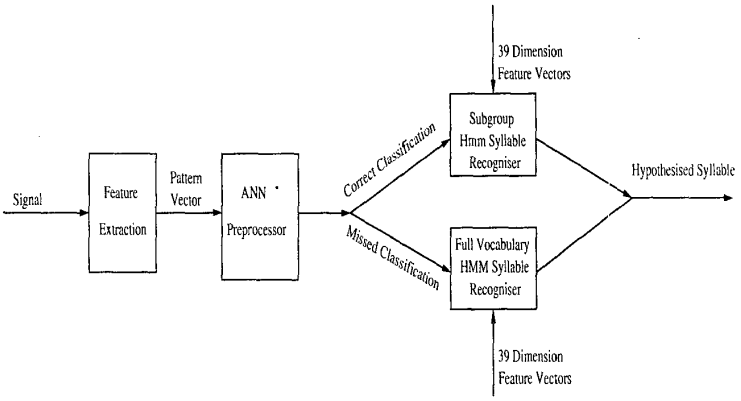

Figure 1. Block diagram of ANN-HMM hybrid system

pattern vectors are expected to capture the dynamics of the syllables. These pattern vectors are used to train a 6 class neural network classifier. The equivalent class to which a syllable maps is determined by replacing all the consonants in it with 'C' and the vowels by 'V'. The resulting syllable structure represents the equivalent class label for the syllable. There are a total of six such classes for the syllable vocabulary chosen. The patterns corresponding to all the cohorts in an equivalent class forms the training set for that class. The neural network classifier is trained to classify the input pattern into one of these six classes. In the testing stage, the output of the preprocessor is the equivalent class label of the syllable being recognised. If the group classification is correct then a HMM based syllable recogniser trained to recognise the cohorts within that class is used for recognising the syllable. For a missed classification an alternative HMM based syllable recogniser developed for the entire syllable vocabulary is used for recognition.

The performance of two hybrid systems, MLFFNN-HMM and a SVM-HMM system using a SVM based preprocessor are compared.

3. RECOGNITION STUDIES

3.1 Database

The task is to recognise syllables in continuous speech in the Telugu language. The Indian language speech corpus described in [8] is used for this study. The database contains read speech by Doordarshan TV news readers of both genders. The speech is segmented and labelled in terms of syllables. Of the 2,273 distinct syllables occurring in the database, we used 268 most frequently occurring syllables as the syllable vocabulary, which covers about 85% of the database. The syllable structure of these syllables were found using the procedure mentioned earlier. These syllables form the following equivalent classses, CCV, CCVC, CV, CVC, V and VC. Among these type of syllable patterns the CV type of syllables are most frequent, followed by the CVC type.

Table 1. Telugu language database

Equivalent class	No. of training examples	No. of testing examples
CCV	2,660	427
CCVC	248	41
CV	44,499	8,333
CVC	9,361	1,774
V	2,348	472
VC	1,031	185
Total	59,547	11,232

The training and test dataset of syllables spliced from continuous speech for the equivalent classes is shown in Table 1.

3.2 MLFFNN Preprocessor

A 6 class MLFFNN was trained using the 130-dimension pattern vectors derived earlier. The patterns corresponding to all the cohorts within one equivalent class form the training set for this classifier. The MLFFNN classifier has a structure 130L 390N 18N 6N, the numbers in the ANN structure represents the number of units in that layer and, L refers to a linear unit and N to a nonlinear unit. The classifier was trained for 200 epochs [9]. In the testing stage a syllable pattern presented to the ANN preprocessor results in an equivalent class label for that pattern. The performance of the ANN preprocessor in classifying the syllables into equivalent classes is shown in Table 2. The classification performance for some of the classes is poor due to the small amount of training data available for them. Also the network is biased towards the CV type of patterns due to the large volume of training examples of that class. In order to minimise the effect of uneven distribution of the training patterns, we explored alternative grouping of the syllable patterns. All syllables containing a stop consonant before the vowel were grouped into one class (CV_1) and the rest of the syllables as another class (CV_2).

Table 2. Performance of the ANN preprocessor

Equivalent class	Recognition rate of individual classes	
	MLFFNN classifier	SVM classifier
CCV	21.08	53.40
CCVC	0	48.78
CV	89.47	94.64
CVC	59.13	69.22
V	53.18	70.76
VC	0	52.43
Overall performance	78.8	87.17

The CVC type of syllables were also similarly grouped. The CCVC type syllables were merged into the CVC_1 or the CVC_2 group. Similarly the CCV group was merged into the CV_1 or CV_2 group depending on the above criteria. A total of seven equivalent classes were formed, namely CV_1, CV_2, CVC_1, CVC_2, V and VC. As another alternative, a smaller classifier of five classes containing only CV_1, CV_2, CVC_1, CVC_2, VVC was developed. Here the VVC group combined all the syllables of type V and type VC together. Both these alternative classifiers performed poorer by 7 to 10% especially for the most frequently occurring CV type of syllables when compared to the original six class group classifier. The overall performance of the hybrid systems using these modified classifiers was about the same. We report results based on the original six class group classifier.

3.3 SVM preprocessor

It would be desirable to achieve a much better performance in the preclassification stage as compared to the MLFFNN preclassifier as this would improve the performance of any cascaded hybrid system. A six class SVM based classifier was trained using the 130-dimension pattern vectors described earlier. One against the rest approach was used for training. The performance of this classifier was much better for all the subgroups. The SVM preclassifier was able to perform better in cases where the MLFFNN classifier failed. The recognition rate for individual classes is shown in Table 2. The overall performance of the preprocessor was better by 9% over that of the MLFFNN preprocessor.

3.4 HMM Based Syllable Recogniser

A HMM based subgroup syllable recogniser was developed for each of the equivalent classes that would recognise the syllables within the cohorts of that class. Another HMM based syllable recogniser was developed for the entire 268 syllable vocabulary. The syllable loop network was used in recognition. All the HMM models were 8 state left to right models and used 2 to 64 mixtures, depending on the available training patterns. From every syllable segment, for a frame size of 15 msec and a frame shift of 5 msec, 13 MFCC along with their delta and acceleration coefficients forming a 39-dimension feature vector were extracted. These features were used for training the HMM models. The models were trained in isolated word fashion. The full vocabulary syllable recogniser was tested using syllable segments spliced from continuous speech. The performance of the recogniser was 56.85%. In earlier studies the best performance for recognition of syllables in continuous speech without segmental information is 52.73% as seen in Table 3 [10].

Table 3. Recognition performance of different systems

Recognition system	Percentage of correct classification
HMM based continuous speech recognition(CSR) system without syllable segment information	52.73
HMM based CSR with syllable segment information	56.85
MLFFNN-HMM hybrid recogniser with syllable segment information	59.48
SVM-HMM hybrid recogniser with syllable segment information	61.58

3.5 MLFFNN-HMM Hybrid System

The results of the preprocessing stage is the class label of the syllable. If the output of the ANN preprocessor is correct, then the subgroup HMM syllable recogniser appropriate for that equivalent class is used to recognise the syllable. On the other hand for a missed preprocessor classification the global HMM syllable recogniser designed for the entire syllable vocabulary is used for recognition. The performance of the hybrid MLFFNN-HMM system is shown in Table 3. The performance of the hybrid system is better than the plain HMM based syllable recogniser in continuous speech (52.73%), where no syllable segment information is used. When syllable segments are tested in isolated word fashion the recognition rate is 56.85% for the HMM syllable recogniser, which is poorer than the hybrid system performance (59.48%).

3.6 SVM-HMM hybrid system

In this system the ANN preprocessor in Figure 1 is the 6 class SVM classifier of Section 3.3. This hybrid system is similar to the MLFFNN-HMM hybrid system. The performance of the SVM-HMM hybrid system is 61.58% an improvement of 5% in absolute terms over that of a HMM only system as seen in Table 3. This improvement is due to the better performance of the preprocessor.

Conventionally the recognition system is considered as a black box, ie., given an input signal the output syllable la-

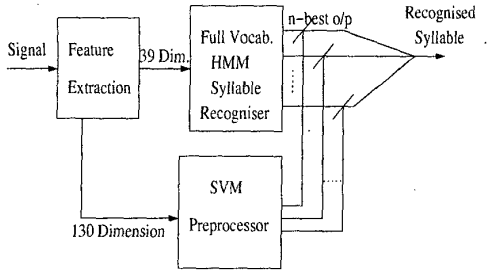

Figure 2. Block diagram of SVM-HMM hybrid system with weighing

bel is desired. In this scenario the SVM preclassifier can be considered as a gating network and each component of the six dimension output of the SVM preclassifier can be treated as a confidence factor for weighting the n-best recognition hypothesis of the HMM syllable recogniser Figure 2 [11]. Depending on the structure of the syllable in each of the 5-best hypothesis of the full vocabulary syllable recogniser, the recognition score is rescored based on the confidence score derived from the SVM preprocessor for that class. The performance of this hybrid system is 54.59%. The performance is similar to the case were no weighting is done (54.44%). In this scenario the weighting of the scores did not help to a large extent, as the 5-best hypothesis of the HMM syllable recogniser belongs more often to the cohorts within a class rather than across classes and hence the output were uniformly weighted. The MLFFNN-HMM hybrid system performance is 48% when it is treated as a black box. It is better than the performance expected (44.79%) when two systems MLFFNN (78.8%) and HMM (56.85%) are cascaded together. Likewise the SVM-HMM hybrid system performance of 54.44% is better than the expected (49.55%), when two systems SVM (87.17%) and HMM (56.85%) are cascaded together.

4. SUMMARY

Two systems, one with an MLFFNN preprocessor and another with a SVM preprocessor were designed. We observe that the SVM preprocessor is better than the MLFFNN preprocessor, especially in categorising the highly confusable CV type of syllables which are the most frequently occurring in a language. The hybrid ANN-HMM systems improve the syllable recognition rate when compared to the best possible HMM based syllable recogniser. We have shown that preprocessing followed by conventional recognition helps in achieving a better recognition rate. In the pattern extraction procedure adopted some frames in the consonant region may be dropped leading to poor representation of the syllable. Alternative strategies based on the vowel onset point [9] may improve the pattern representation and consequently the performance of the preprocessor. The superior performance of the SVM based preprocessor with a better discriminating HMM models at the decision stage may help achieve a further improvement. We have also show that the output of the SVM can be used as a weight metric for enhancing the evidence from the subsequent recogniser. It will be especially useful when the subsequent recogniser has more inter class confusion.

REFERENCES

[1] P. Eswar, S. K. Gupta, C. Chandra Sekhar, B. Yegnanarayana, and K. Nagamma Reddy, "An acoustic-phonetic expert for analysis and processing of continuous speech in Hindi," in Proc. European Conf. Speech Technology, (Edinburgh, UK), pp. 369–372, Sept. 1987.

[2] Suryakanth V. Gangashetty, K. Sreenivasa Rao, A. Nayeemulla Khan, C. Chandra Sekhar, and B. Yegnanarayana, "Combining evidence from multiple modular networks for recognition of consonant-vowel units of speech," in Proc. IEEE Int. Joint Conf. Neural Networks, vol. 1, (Portland, Oregon, USA), pp. 686–691, July 2003.

[3] C. Chandra Sekhar and B. Yegnanarayana, "A Constraint satisfaction model for recognition of Stop Consonant-Vowel (SCV) utterances," IEEE Trans. Speech Audio Processing, vol. 10, pp. 472–480, Oct. 2002.

[4] Teuvo Kohonen, Self-Organising Maps, vol. Springer Series in Information Sciences, Vol. 30. 2001.

[5] Harvé Bourlard, Connectionist Speech Recognition - A Hybrid Approach. Boston, USA: Kluwer Academic Publishers, 1994.

[6] Lawrence. R. Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition. New Jersy: PTR Prentice Hall Inc, 1993.

[7] C. Chandra Sekhar, W. F. Lee, K. Takeda, and F. Itakura, "Acoustic modeling of subword units using support vector machines," in Workshop on Spoken Language Processing, (TIFR, Mumbai, India), pp. 79–86, Jan. 2003.

[8] A. Nayeemulla Khan, Suryakanth V. Gangashetty, and S. Rajendran, "Speech database for Indian languages-A preliminary study," in Proc. Int. Conf. Natural Language Processing, (NCST, Mumbai, India), pp. 295–301, Dec. 2002.

[9] Suryakanth V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Dimension reduction using autoassociative neural network models for recognition of consonant-vowel units of speech," in Fifth Int. Conf. Advances in Pattern Recognition. (To be presented).

[10] A. Nayeemulla Khan, "Recognition of syllable like units in Indian languages," in Proc. Int. Conf. Natural Language Processing (T. be presented, ed.).

[11] Simon Haykin, Neural Networks a Comprehensive Foundation. Pearson Education Inc., 1999.