Proceedings of
ICISIP - 2005

Combining Evidence from Multiple Classifiers for Recognition of Consonant-Vowel Units of Speech in Multiple Languages

Suryakanth V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
{svg,chandra,yegna}@cs.iitm.ernet.in

Abstract

In this paper, we present studies on combining evidence from multiple classifiers to recognize a large number of consonant-vowel (CV) units of speech. Multiple classifier systems may lead to a better solution to the complex speech recognition tasks, when the evidence obtained from individual systems is complementary in nature. Hidden Markov models (HMMs) are based on the maximum likelihood (ML) approach for training CV patterns of variable length. Support vector machine (SVM) models are based on discriminative learning approach for training fixed length CV patterns. Because of the differences in the training methods and in the pattern representation used, they may provide complementary evidence for CV classes. Complementary evidence available from these classifiers is combined using the sum rule. Effectiveness of the multiple classifier system is demonstrated for recognition of CV units of speech in Indian languages.

1. INTRODUCTION

For any pattern recognition task, there are a number of approaches to the design of classification models based on different principles of learning from examples. For complex recognition tasks involving a large number of classes and noisy data, it is difficult to achieve desirable performance using a single system. Features and classifiers of different types complement one another [1]. Therefore features and classifiers of different types can be combined to improve the recognition performance. Hidden Markov models (HMMs) used in speech recognition are based on the maximum likelihood (ML) approach for training. The incremental model optimization approach in ML framework simplifies the training process, but does not use discriminative learning [2]. This is due to the fact that training data corresponding to other models are not considered during the optimization of parameters for a given model. Training by optimization over the entire pattern space gives better discriminative power to the models since the models learn patterns that need to be discriminated. Support vector machines (SVMs) are good at this type of learning since the training involves optimization over the entire pattern space [3]. The SVMs have attained prominence due to their inherent discriminative learning and generalization capabilities from the limited training data. These models learn the boundary regions between patterns belonging to two classes by mapping the input patterns into a higher dimensional space, and seeking a separating hyperplane so as to maximize its distance from the closest training examples. The choice of kernel function best-suited for mapping the input patterns into a higher dimensional space for a given task is an open problem.

The evidence obtained from SVM and HMM classifiers could be complementary in nature. The outputs of these classifiers may be interpreted as the evidence available for classes. Combination of these evidence is expected to result in an improved performance. In this study, we combine the evidence from SVM and HMM classifiers using the sum rule to perform recognition of consonant-vowel (CV) units of speech [4].

The paper is organized as follows: In Section 2, we describe the proposed method for integration of multiple classifiers for recognition of CV units. Section 3 demonstrates the enhancement of evidence using the integrated system. In Section 4, we present the studies on recognition of CV units of speech in Indian languages.

2. INTEGRATION OF MULTIPLE CLASSIFIERS FOR RECOGNITION OF CV UNITS

Acoustic modeling of context dependent CV units of speech using SVM models involves training classifiers with pattern vectors of fixed dimension extracted from CV segments of varying durations. To derive patterns of fixed dimension, the instant at which the consonant ends and vowel begins in a CV utterance, called the vowel onset point (VOP), is detected. A segment of fixed duration around the VOP contains most of the information necessary for recognition of CV units. This segment is processed to derive fixed dimension patterns automatically from varying duration CV segments. Portions of a CV utterance in the beginning and end are not included in the fixed duration segment, since they may be affected by the coarticulation effects. This type of pattern representation may lead to loss of some information.

HMM based systems are capable of handling varying length patterns. All frames of a CV utterance are used for processing. In this type of pattern representation, portions of CV utterance in the beginning and the end are also included. These portions may be affected by the coarticulation effects. This may lead to confusion among the sound units. For estimation of HMM parameters, it is necessary to use a large training data set. Due to varying frequency of occurrence of different CV classes it is difficult to collect sufficiently large number of examples for many classes.

The performance of the SVM and HMM based systems is expected to be different because of the difference in the training methods and in the pattern representations used. Therefore, we propose to integrate these two types of classifiers to obtain an improved performance. Now, we discuss the design and implementation of individual systems and their integration.

2.1. SVM based Recognition System

Once the VOP in a CV segment is hypothesised, a 65 msec segment around the VOP is processed using short-time analysis of speech signal to obtain a sequence of spectral feature vectors. Five overlapping frames to the left of VOP and five to the right of VOP are included in the fixed length pattern representation of the CV segment [5]. A frame shift of 5 msec and a frame size of 20 msec are used. Each frame is represented by a 39-dimension feature vector consisting of 12 mel-frequency cepstral coefficients (MFCC), energy, their first order derivatives (delta coefficients) and their second order derivatives (acceleration coefficients). A fixed dimension pattern vector of 390-dimension is obtained by concatenating the feature vectors of 10 successive frames. Since the dimension of CV pattern vector is large, nonlinear compression using an autoassociative neural network (AANN) models have been considered to obtain a reduced dimension pattern vector without significant loss of information [5][6].

The block diagram of the system using SVM models for recognition of CV units is shown in Fig. 1. It consists of three stages. In the first stage, the 390-dimensional input pattern vector $\mathbf{x}$ is compressed to 60-dimensional vector using an AANN with structure $390L$ $585N$ $60N$ $585N$ $390L$, where L refers to linear units and N refers to nonlinear units [5]. These compressed pattern vectors are used to train the SVM models. One-against-the-rest approach is used for decomposition of the learning problem from a n-class pattern recognition into several two-class learning problems. An SVM is constructed for each class by discriminating that class against the remaining $(n-1)$ classes. The recognition system based on this approach consists of n SVM models. The set of training examples $\{\{(\mathbf{x}_i, k)\}_{i=1}^{N_k}\}_{k=1}^{n}$ consists of N_k number of examples belonging to k^{th} class, where the class label $k \in \{1, 2, \ldots, n\}$. All the training examples are used to construct an SVM model for a class. The SVM model for the class k is constructed using a set of training examples and their desired outputs, $\{\{(\mathbf{x}_i, y_i)\}_{i=1}^{N_k}\}_{k=1}^{n}$. The desired output y_i for a training example $\mathbf{x}_i$ is defined as follows:

$$y_i(\mathbf{x}_i) = \begin{cases} +1 & : \ \mathbf{x}_i \ belongs\ to\ k^{th}\ class \\ -1 & : \ otherwise \end{cases}$$

For a given test pattern $\mathbf{x}$, the evidence $D_k(\mathbf{x})$ is obtained from each of the SVMs. The normalised evidence $S_k(\mathbf{x})$ in the range of 0 to 1, is used for further processing.

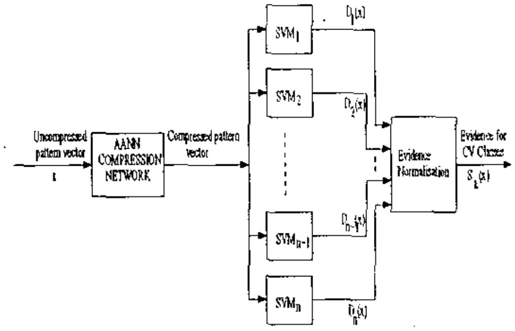

Figure 1. Block diagram of the CV recognition system using SVM models.

2.2. HMM based Recognition System

The block diagram of the system using HMM models for recognition of CV units is shown in Fig. 2. A CV segment is analyzed frame by frame, with a frame size of 20 msec and a shift of 5 msec. Each frame is represented by a 39-dimension MFCC based feature vector as explained as earlier. All the frames of a CV segment are used in the pattern representation of the CV segments. A 5-state, left-to-right, continuous density HMM using multiple mixtures with diagonal covariance matrix is trained for each class. For the CV classes with a frequency of occurrence less than 100 in the training data set, 2 mixtures are chosen. The number of mixtures is 4 for those CV classes whose frequency of occurrence is between 100 and 500. For the other classes, the number of mixtures is 8. For the observation sequence x of a test utterance, the likelihood $P(\mathbf{x}|M_i)$ is obtained from each of the HMMs (M_i). The log-likelihood values are scaled to the range of 0 to 1. The normalised evidence $H_k(\mathbf{x})$ is used for further processing.

2.3 Integration of SVM and HMM based Systems

The SVM and HMM classifiers are trained independently. The evidence is obtained for each CV unit from the outputs of the two classifiers. The block diagram of the proposed system built by integrating SVM and HMM classifiers is shown in Fig. 3. The sum rule is considered for combination as it has been shown to be effective in combining the complementary evidence from multiple classifiers [4]. In the decision logic,

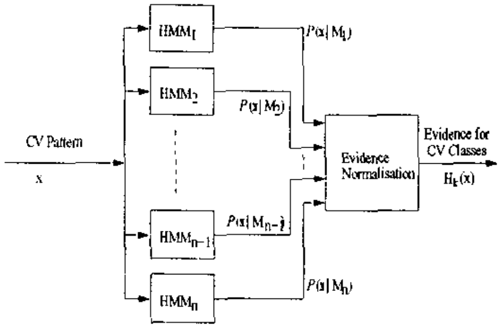

Figure 2. Block diagram of the CV recognition system using HMM models.

the class label associated with the highest combined evidence is hypothesised as the class of the test pattern. That is,

$$Class(\mathbf{x}) = \max_{k} \quad (S_k(\mathbf{x}) + H_k(\mathbf{x})) \qquad (1)$$

where $S_k(\mathbf{x})$ and $H_k(\mathbf{x})$ corresponds to the normalised evidence from the k^{th} SVM and HMM models respectively, for the input test utterance x.

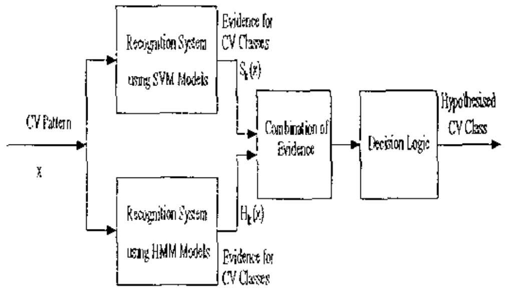

Figure 3. Block diagram of the proposed CV recognition system built by integration of SVM and HMM based classifiers.

3. ENHANCEMENT OF EVIDENCE

Enhancement of evidence by the integrated system is demonstrated for a test utterance of the CV class /du/. The index of the class /du/ is 1. The outputs of the SVM models are shown in Fig. 4(a). The first five class indices in decreasing order of output (evidence) values are {2, 9, 1, 135, 12}, corresponding to classes {/di/, /da/, /du/, /dhi/, /ta/}. Similarly the outputs of the HMM models are shown in Fig. 4(b). The first five class indices in decreasing order of output values are {65, 16, 1, 3, 9}, corresponding to classes {/Nu/, /Du/, /du/, /ha/, /da/}. It is seen that the output for /du/ is not the largest for both types of classifier systems. Therefore, both the systems would have misclassified the test utterance. Though the evidence for /du/ is not the highest in both the cases, it is significantly high. The combined evidence determined using the outputs of SVM and HMM classifiers is plotted in Fig. 4(c). In this case, the first five class indices in decreasing order of combined confidence values are {1, 9, 2, 3, 135}, corresponding to classes {/du/, /da/, /di/, /ha/, /dhi/}. It is seen that the class /du/ has the largest value indicating that the combined evidence is the highest for that CV class. This behavior of the integrated system to enhance the evidence for a CV class based on the evidence from multiple systems is helpful in improving the recognition performance.

4. STUDIES ON RECOGNITION OF CV UNITS

In this section, we present our study on recognition of CV units of speech in three Indian languages. Speech database consisting of recordings of TV news bulletins in Tamil, Telugu and Hindi languages is used in our studies. A brief description of the speech corpus used in our studies is given in Table 1. Each bulletin contains 10 to 15 minutes of speech from a single speaker. The CV units in the database are segmented and labeled manually. These units have varying frequencies of occurrence in the database. We consider a set of 196 CV classes that occur more than 40 times in the training data.

We first study the recognition performance of the SVM based system. The k-best recognition performance for 196 CV classes is given in Table 2. The $k-$best performance corresponds to the case when the actual class of a test pattern is present among top k classes with the largest valued evidence. The recognition performance of CV units for the HMM based system is also given in Table 2. It is seen from Table 2 that the SVM based system performs better than that based on HMMs. The SVMs use discriminative information in the process of learning, whereas HMM models are trained using ML framework which does not use discriminative learning [2].

Table 2 gives the performance of the integrated system that combines evidence from the SVM and HMM based systems using sum rule. It is seen that the system based on combination of evidence leads to a marginal increase in the classification accuracy in comparison with the best individual system (48.72% versus 45.31%).

The performance of different classifiers for each of the classes is given in Fig. 5. Each entry in the figure shows the percentage of the total number of test patterns of a CV class that are correctly classified by a particular classification system. It can be seen from Figs. 5(a) and (b) that SVM and HMM based systems do not give the same performance for many classes. After combining evidence from these two systems, there is a marginal improvement in the performance for many classes, as seen in Fig. 5(c).

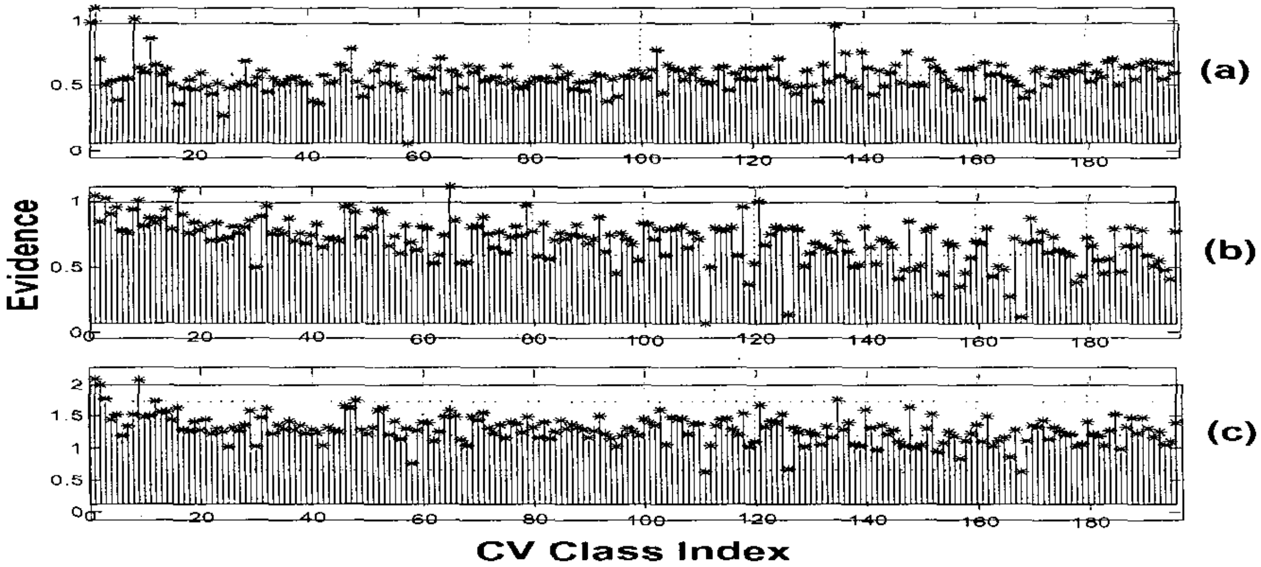

Figure 4. Evidence for CV units from (a) SVM based system, (b) HMM based system and (c) Multiple classifier system for an utterance of class /du/.

5. SUMMARY AND CONCLUSIONS

In this paper, we propose an approach based on a combination of multiple classifier systems for recognition of consonant-vowel (CV) units of speech in multiple languages. Classifiers based on support vector machines and hidden Markov models are designed. These two types of systems use different learning methods and pattern representation. Hence, they may provide complementary evidence for CV classes. The evidence from the two classifier systems is combined using the sum rule. Effectiveness of the proposed approach is demonstrated for recognition of CV units of speech in Indian languages. The combination of evidence leads to a marginal increase in the classification performance.

REFERENCES

[1] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," IEEE AaTrans.nd Machine Intelligence , vol. 16, no. 1, pp. 66–75, Jan. 1994.

[2] H. Bourlard and N. Morgan, Connectionist Speech Recognition: A HApbriodach , Kluwer Academic Publishers, Boston, 1994.

[3] A. J. Robinson, DyPnraompiacgaEtiorrnor Networks, PhD thesis, Engineering Department, Cambridge University, Feb. 1989.

[4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers,PIaEtEtEern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[5] S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Dimension reduction using autoassociative neural network models for recognition of consonant-vowel units of speech," in Proc. fth Int. ConEf.ngoinneAerdinvgan(cIeCsAiEnI Calcutta, India), Dec. 2003, pp. 156–159.

[6] K. I. Diamantaras and S. Y. Kung, Principal TeCNcoehmtnwpioqonurkeesn,t Applications, John Wiley and Sons, Inc., New York, 1996.

Table 1. Description of broadcast news speech corpus used in studies.

Number of bulletins (Tamil:Telugu:Hindi)	72 (33:20:19)
Gender of news readers (Male:Female)	(27:45)
Number of bulletins used for training (Male:Female)	59 (22:37)
Number of bulletins used for testing (Male:Female)	13 (5:8)
Number of CV classes used for the study	196
Number of CV segments used for training	1,05,502
Range of frequency of occurrence for the classes in the training data	40 to 2,826
Number of CV segments used for testing	25,777

Table 2. Comparison of the $k-$best classification performance for different CV recognition systems.

CV recognition system	$k-$best classification performance (in %)				
	k=1	k=2	k=3	k=4	k=5
SVM based system	45.31	57.62	64.00	68.08	71.03
HMM based system	41.32	47.46	50.80	52.91	54.57
Integration of SVM and HMM based classifiers using sum rule	48.72	62.55	69.73	74.12	77.32

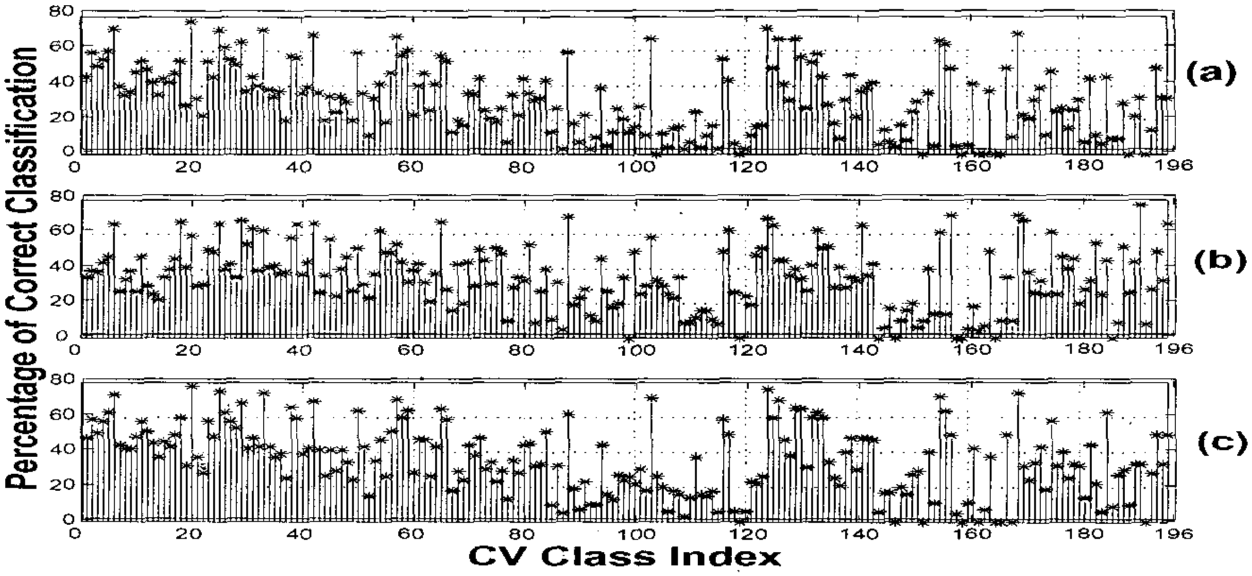

Figure 5. Classification performance (in percentage) of recognition systems based on (a) SVMs, (b) HMMs and (c) Integration of SVM and HMM systems, for the test data of each CV class.