Acoustic Model Combination for Recognition of Speech in Multiple Languages using Support Vector Machines

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
email: {svg,chandra,yegna}@cs.iitm.ernet.in

Abstract—**In this paper, we study the performance of support vector machine based classifiers in acoustic model combination for recognition of context dependent subword units of speech in multiple languages. In acoustic model combination, the data for similar subword units across languages are shared to train acoustic models for multilingual speech. Sharing of data across languages leads to an increase in the number of training examples for a subword unit common to the languages. It may also lead to increase in the variability of the data for a subword unit. In this paper, we study the effect of data sharing on the classification accuracy and complexity of acoustic models built using support vector machines. We compare the performance of multilingual acoustic models with that of monolingual acoustic models in the recognition of a large number of consonant-vowel units in the broadcast news corpus of three Indian languages.**

I. Introduction

Development of acoustic models for multilingual subword units is important for recognition of speech in multiple languages. Acoustic model combination to share acoustic models for similar subword units across languages has been proposed in [1]. In the acoustic model combination method, the set of subword units for multilingual speech recognition system is derived by forming a union of sets of units in different languages. Data available for a subword unit common to multiple languages can be shared to train an acoustic model for that unit. Sharing of data across languages leads to availability of larger training data sets and increased variability in the data set. In this paper, we study the effect of data sharing on multilingual acoustic models built using support vector machine (SVM) models.

Multiclass pattern recognition systems using SVMs are commonly built using the one-against-the-rest approach [2]. In this approach, one SVM model is built for each class. For each SVM model an optimal hyperplane is constructed in the kernel feature space to separate the examples of a class from the examples of all the other classes. Data sharing in acoustic model combination leads to an increase in the number of negative examples for all the classes. However, the number of positive examples will increase only for the classes that are common to multiple languages. Nonuniform increase in the number of positive examples for different classes affects the classification accuracy and complexity of SVMs. In this paper, we study the effect of data sharing on the performance and complexity of multilingual acoustic models for recognition of consonant-vowel (CV) utterances of speech in three Indian languages.

The paper is organized as follows: In the next section, we describe the multiclass pattern recognition system for recognition of CV units of speech using SVMs. In Section III, we present the studies on recognition of monolingual and multilingual CV units. In Section IV, we present the effect of data sharing on the complexity of SVMs.

II. Consonant-vowel Recognition System using SVM Models

Most production and perception units of speech sounds are dynamic in nature. Due to existence of coarticulation effects of sound units in syllable-like units, they are chosen as subword units. Pronunciation variation is more systematic at the level of syllables compared to the phoneme level. Syllable-like units such as consonant-vowel (CV) units are important information-bearing sound units from production and perception point of view [3]. Therefore, we considered CV units of speech as the basic units for speech recognition.

SVM classifiers require a fixed length pattern representing the CV utterances. For this, the instant at which the consonant ends and the vowel begins in a CV utterance, called the vowel onset point (VOP), is detected. The approach used for detection of VOP is based on dynamic time alignment between a reference pattern of a CV class and any other pattern representing that class [4]. Once the VOP is hypothesised, five overlapping frames are considered to the left of VOP and five to the right, with a shift of 5 msec. The duration of each frame is 20 msec. Each frame is represented by a feature vector consisting of 12 mel-frequency cepstral coefficients (MFCC), energy, their first order derivatives (delta coefficients) and their second order derivatives (acceleration coefficients). Dynamic features such as delta and acceleration coefficients are used to capture the change in the characteristics of the speech signal for dynamic sound units such as CV utterances [5][6]. The dimension of each frame is 39. Thus, the pattern vector formed

0-7803-8359-1/04/$20.00 ©2004 IEEE

TABLE I

Description of broadcast news speech corpus used in studies on CV recognition.

Description	Language			
	Monolingual			Multilingual
	Tamil	Telugu	Hindi	
Number of bulletins	33	20	19	72
News readers (Male:Female)	(10:23)	(11:9)	(6:13)	(27:45)
Number of bulletins used for training (Male:Female)	27 (8:19)	16 (9:7)	16 (5:11)	59 (22:37)
Number of bulletins used for testing (Male:Female)	6 (2:4)	4 (2:2)	3 (1:2)	13 (5:8)
Number of CV classes used for the study	123	138	103	196
Number of CV segments used for training	43,541	41,725	20,236	1,05,502
Range of frequency of occurrence for the classes in the training data	39 to 1,633	40 to 2,037	40 to 1,264	40 to 2,826
Number of CV segments used for testing	10,293	11,347	4,137	25,777

for each CV utterance is a 390-dimension vector formed by concatenating the feature vectors of 10 successive frames. To reduce computation complexity, we propose nonlinear compression of the large dimension input pattern vectors using autoassociative neural network (AANN) models [7][8].

The block diagram of the proposed system for recognition of CV units is shown in Fig. 1. It consists of three stages. In the first stage, the 390-dimension input pattern vectors $\mathbf{x}$ are compressed to 60-dimension, using an AANN with structure *390L 585N 60N 585N 390L*, where L refers to linear units and N refers to nonlinear units [7]. Further, these compressed pattern vectors are used to train the SVM classifier. One-against-the-rest approach is used for decomposition of the learning problem in 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 number of SVMs. 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 for a class. The SVM 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 examples with $y_i = +1$ are called positive examples, and those with $y_i = -1$ are called negative examples. An optimal hyperplane is constructed to separate positive examples from negative examples. The separating hyperplane (margin) is chosen in such a way as to maximize its distance from the closest training examples of different classes [9]. The support vectors are those data points that lie closest to the decision surface, and therefore the most difficult to classify. They have a direct bearing on the optimum location of the decision surface. For a given test pattern $\mathbf{x}$, the evidence $D_k(\mathbf{x})$ is obtained from each of the SVMs. In the decision logic, the class label k associated with the SVM which gives maximum evidence is hypothesised as the class of the test pattern.

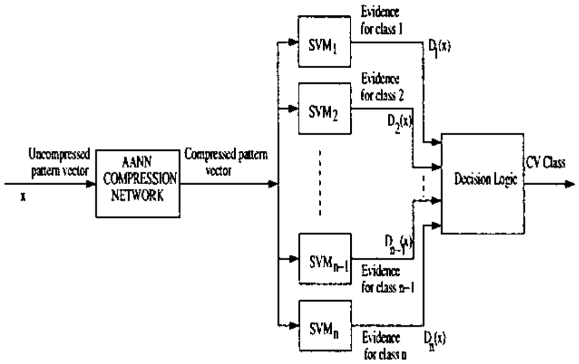

Fig. 1. Block diagram of the CV recognition system using AANN model for compression of pattern vectors followed by SVM models for classification.

The complexity of the SVM model for a class depends on the number of positive examples, the number of negative examples and the margin of separation between them in the kernel feature space. Data sharing in acoustic model combination leads to an increase in the number of negative examples for all the classes. The increase in the number of positive examples for a class depends whether it is common to multiple languages. For classes present in only one language, there will not be any increase in the number of positive examples. Therefore, the effect of data sharing may be nonuniform for different classes. In the following sections, we compare the performance and complexity of multilingual acoustic models with that of the monolingual models.

III. Recognition of Monolingual and Multilingual CV Units

Speech database consisting of recordings of TV news bulletins in Tamil, Telugu and Hindi languages is used in our studies. A summary of the database for these three languages is given in Table I. 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 CV classes that occur more than 50 times in the database.

The recognition systems are developed separately for each of the three languages. The k-best recognition performance for 123 CV classes in Tamil, 138 classes in Telugu and 103 classes in Hindi languages is given in Table II. For comparison, the performance of the HMM based systems is obtained. The HMM based system is a 5-state, left-to-right, continuous density multiple mixtures with diagonal covariance matrix and it is trained separately for each class. The number of mixtures is 2 for the CV classes with a frequency of occurrence less than 100 in the training data. The number of mixtures is 4 for those CV classes whose frequency of occurrence is in between 100 and 500. For the other classes, the number of mixtures is 8. Each frame is represented by a 39-dimension parametric vector as explained earlier. The recognition performance of CV units for each of the three languages using HMM models is also given in Table II.

TABLE II
COMPARISON OF THE k-BEST CLASSIFICATION PERFORMANCE FOR DIFFERENT MONOLINGUAL CV RECOGNITION SYSTEMS.

Language	System	$k-$best classification performance				
		k=1	k=2	k=3	k=4	k=5
Tamil	HMM	50.55	57.28	60.72	62.67	64.24
	SVM	50.18	64.00	70.33	74.66	77.26
Telugu	HMM	46.48	53.05	56.58	58.85	60.55
	SVM	50.61	62.57	68.52	72.45	75.42
Hindi	HMM	40.09	46.28	49.13	51.06	52.93
	SVM	41.04	52.91	59.92	65.02	67.99

It is seen from Table II that the 5-best performance obtained using SVM classifiers is significantly better than HMM based system for all the three languages. The improved performance is mainly due to the fact that SVM uses discriminative information in the process of training. The 1-best performance of systems based on both types (HMM and SVM) of models performed is nearly the same for all the three languages.

Next, we describe a multilingual system in which the acoustic models for similar CV classes across languages are combined. The similar classes from different languages are derived from Indian Language Transliteration (ITRANS) code [10]. The ITRANS code was chosen, as it uses the same symbol to represent the sound across the Indian languages. A summary of the database used for the development of multilingual system is given in the last column of Table I. After combining 123, 138, and 103 classes from Tamil, Telugu and Hindi respectively, we get 196 unique classes.

SVM models are generated by assigning one model to each CV class, and training this model by considering data from all the three languages. The k-best recognition performance for 196 CV classes is given in Table III. For comparison, the performance of the HMM based system is obtained for the 196 classes. The recognition performance of CV units for HMM system is also given in Table III.

TABLE III
COMPARISON OF THE k-BEST CLASSIFICATION PERFORMANCE FOR MULTILINGUAL CV RECOGNITION SYSTEMS.

System	$k-$best classification performance				
	k=1	k=2	k=3	k=4	k=5
HMM	41.32	47.46	50.80	52.91	54.57
SVM	45.31	57.62	64.00	68.08	71.03

It is seen from Table III that the SVM based multilingual system performs significantly better than that based on HMMs. The difference in the performance is more significant compared to the monolingual systems. SVMs use discriminative information in the process of learning, whereas HMM models are trained using ML framework which loses discriminative information [11]. Due to this fact there exists significant difference (71.03% versus 54.57%) in the 5-best classification performance.

The classification performance of the multilingual system for the CV classes in different languages is tabulated separately in Table IV. The classification performance of the monolingual systems is copied from Table II for comparison. The performance of HMM based multilingual systems for each of the languages is less by 3 to 6%, whereas for SVM based multilingual systems it is less by 2 to 5%.

TABLE IV
COMPARISON OF THE 1-BEST CLASSIFICATION PERFORMANCE FOR MONOLINGUAL AND MULTILINGUAL CV RECOGNITION SYSTEMS.

Language	System	1-best classification performance	
		Monolingual	Multilingual
Tamil	HMM	50.55	44.06
	SVM	50.18	45.69
Telugu	HMM	46.48	40.38
	SVM	50.61	47.38
Hindi	HMM	40.09	37.09
	SVM	41.04	38.72

In Table V, we give the average performance of the systems. Though the SVM based multilingual system is discriminating larger number (196) of classes across three languages, its performance is only marginally less compared to the average performance of the monolingual systems. Also, HMM based system classification performance is affected by the large number of classes. These observations indicate that when SVM models are used for classification, the number of classes has less effect on the recognition performance. The classification performance of most of the classes is slightly less when the classes are combined from the three languages in comparison with the monolingual systems.

TABLE V

COMPARISON OF MULTILINGUAL SYSTEM WITH AVERAGE PERFORMANCE OF MONOLINGUAL SYSTEMS FOR k=1.

System	Classification performance	
	Average of monolinguals	Multilingual
HMM	47.07	41.32
SVM	48.90	45.31

IV. Effect of Data Sharing on Complexity of SVM Models

The percentage of increase (or decrease) in support vectors $S(k)$ for each of the classes in multilingual system in comparison with the union of monolingual systems is given by:

$$S(k) = \frac{N_m(k) - (N_{Ta} + N_{Te} + N_{Hi})}{(N_{Ta} + N_{Te} + N_{Hi})} \times 100$$

where $N_m(k)$ is the number of support vectors (SV) used by a class k in multilingual system and N_{Ta}, N_{Te}, N_{Hi} are the number of SVs used by the corresponding classes in Tamil, Telugu and Hindi monolingual systems, respectively.

The measure $S(k)$ obtained for each of the 196 classes is plotted in Fig. 2. The number of classes are grouped into three subgroups. Classes occurring in all the three languages, occurring in any two languages and occurring in only one of the three languages. There are 52 classes that belong to first category, 64 to second category and 80 to third category. The $S(k)$ values are plotted in this order. For those classes that occur in all the three languages, there is a decrease in percentage of SVs. For those classes present in only one of the three languages, the increase in percentage of SVs is more. The percentage of increase is slightly less for the classes which occur in any two languages.

The complexity of monolingual and multilingual systems in terms of average number of support vectors per class is examined. It is found that the number of support vectors are 3,112 and 3,203 for monolingual and multilingual systems, respectively. The complexity of these two systems is nearly the same. This observation suggests that whether the multilingual system is designed using a set of monolingual systems or as a combination of similar classes from different languages, the complexity remains nearly the same.

V. Summary and Conclusions

In this paper, we proposed a recognition system for consonant-vowel (CV) utterances of speech in three Indian languages. The acoustic models for similar CV classes across the languages are combined. Our approach is motivated by the common characteristics among many sound units across the Indian languages. Support vector machines (SVM) are used for classification. The large dimension input pattern vectors are compressed using autoassociative neural network models. The results of our studies show that the SVM based recognition system gives better performance compared to the systems based on hidden Markov models. This is due to the discriminative information used in the training process. Though the variability among the data set and and number of classes is more for the multilingual system, it has less effect on the recognition performance when SVMs are used for classification. The complexity of multilingual system and monolingual systems is nearly the same, in terms of the number of support vectors.

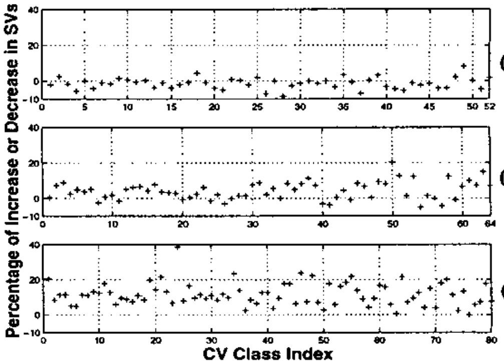

Fig. 2. Percentage of increase (decrease) in number of support vectors for classes of multilingual system in comparison with the combined SVs for the classes present in (a) all the three, (b) any two and (c) only one of the three languages.

References

[1] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, Aug. 2000.

[2] C. Chandra Sekhar, K. Takeda, and F. Itakura, "Recognition of consonant-vowel (CV) units of speech in a broadcast news corpus using support vector machines," in *Proc. Int. Workshop on Pattern Recognition using Support Vector Machines (Niagara Falls, Canada)*, Aug. 2002, pp. 171–185.

[3] 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*, Sep. 1987, pp. 369–372.

[4] S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Extraction of fixed dimension patterns from varying duration segments of consonant-vowel utterances," in *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing (Chennai, India)*, Jan. 2004, pp. 159–164.

[5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[6] L. R. Rabiner and B. -H. Juang, *Fundamentals of Speech Recognition*, PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[7] 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. Fifth Int. Conf. Advances in Pattern Recognition (ISI Calcutta, India)*, Dec. 2003, pp. 156–159.

[8] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks, Theory and Applications*, John Wiley and Sons, Inc., New York, 1996.

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall International, New Jersey, 1999.

[10] A. Chopde, "ITRANS Indian Language Transliteration Package Version 5.2," *Source, http://www.aczone.com/itrans/.*

[11] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, Boston, 1994.