Spotting Multilingual Consonant-Vowel Units of Speech Using Neural Network Models

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. Multilingual speech recognition system is required for tasks that use several languages in one speech recognition application. In this paper, we propose an approach for multilingual speech recognition by spotting consonant-vowel (CV) units. The important features of spotting approach are that there is no need for automatic segmentation of speech and it is not necessary to use models for higher level units to recognise the CV units. The main issues in spotting multilingual CV units are the location of anchor points and labeling the regions around these anchor points using suitable classifiers. The vowel onset points (VOPs) have been used as anchor points. The distribution capturing ability of autoassociative neural network (AANN) models is explored for detection of VOPs in continuous speech. We explore classification models such as support vector machines (SVMs) which are capable of discriminating confusable classes of CV units and generalisation from limited amount of training data. The data for similar CV units across languages are shared to train the classifiers for recognition of CV units of speech in multiple languages. We study the spotting approach for recognition of a large number of CV units in the broadcast news corpus of three Indian languages.

1 Introduction

The main objective of continuous speech recognition system is to provide an efficient and accurate mechanism to transcribe human speech into text. Typically, continuous speech recognition is performed in the following two steps: (1) speech signal to symbol (phonetic) transformation, and (2) symbol to text conversion. Two approaches are commonly used for subword unit based continuous speech recognition. The first approach is based on segmentation and labelling [1]. In this approach, the continuous speech signal is segmented into subword unit regions and a label is assigned to each segment using a subword unit classifier. The main limitation of this approach is the difficulty in automatic segmentation of continuous speech into subword unit regions of varying durations. Because of imprecise articulation and coarticulation effects, the segment boundaries are manifested poorly. The second approach to speech recognition is based on building word

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 303–317, 2005.

models as compositions of subword unit models, and recognising sentences by performing word-level matching and sentence level matching using word models and language models respectively [1]. The focus of this approach is on recognising higher level units of speech such as words and sentences rather than on recognising subword units.

In this paper, we propose an approach for multilingual speech recognition by spotting subword units. Specifically, we consider a method for spotting subword units using vowel onset points (VOPs) as anchor points and labelling the regions around these VOPs using suitable classifiers. The important features of spotting approach are that there is no need for automatic segmentation of speech and it is not necessary to use models for higher level units to recognise the subword units. The symbols that capture the phonetic variations of sounds are suitable units for signal to symbol transformation. 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 [2]. Therefore, we consider CV units of speech as the basic subword units for speech recognition. In Indian languages, the CV units occur with high frequency.

The distribution capturing ability of autoassociative neural network (AANN) models is explored for detection of VOPs in continuous speech [3]. An important issue for the development of a suitable classification system for the recognition of CV units in Indian languages is the large number of these units. Combination of more than 30 consonants and 10 vowels of a language result in a set of about 300 CV units. Further, there are many regional languages across the country. Difficulties in the development of multilingual speech recognition systems are due to the presence of several new classes, degree of overlapping of classes and frequency of occurrence of a given class in different languages. The difficulties in designing a multilingual system are also due to variability among the data set, amount of training data and large number of CV classes. Also, many of the CV units have similar acoustic features. Additionally, the number of examples available in a corpus is not the same for all the units. There may be many units for which only a small number of examples are available. We consider a data sharing approach for development of classification system by combining same type of CV classes across the Indian languages [4]. We consider support vector machine (SVM) based classifiers due to their ability of generalization from limited training data and also due to their inherent discriminative learning [5]. The variability among the data set and more number of classes in multiple languages has less effect on the recognition performance when SVMs are used for classification [4]. However, the application of SVMs to speech recognition problems has been limited to smaller vocabulary tasks due to computational complexity. To reduce the computational complexity, we propose nonlinear compression of large dimensional input pattern vectors using the dimension reduction capability of autoassociative neural network models [6] [7]. We demonstrate the CV spotting based approach to continuous speech recognition for sentences in multiple Indian languages.

The paper is organised as follows: In Section 2, we discuss the issues in spotting CV units. The description of speech corpus used in the studies is given in Section 3. Studies on detection of VOPs in continuous speech utterances are described in Section 4. In Section 5, we present the studies on recognition of multilingual CV units. The system for spotting multilingual CV units in continuous speech is described in Section 6. In this section the spotting approach is illustrated with an example. Studies on recognition of CV units by processing the segments around the hypothesised VOPs in continuous speech utterances are also presented in this section.

2 Issues in Spotting Multilingual CV Units

Strategies for spotting subword units in continuous speech have been based on training the classifiers to recognise only the segments of the continuous speech signal belonging to subword units and reject all other segments. The models thus trained to classify or reject are then used to scan the speech signal continuously and hypothesise the presence or absence of the corresponding subword units. This strategy is similar to the keyword spotting approaches [8]. The main limitation of this strategy based on scanning is that a large number of spurious hypotheses are given by the spotting system [9]. For spotting CV units in continuous speech, we consider an approach based on detection of VOPs and labelling the segments around the VOPs using SVM based CV classifier [4] [10]. The main issues in spotting CV units in the proposed approach are development of a method for detection of VOPs with good accuracy and development of an SVM based classifier capable of discriminating large number of CV classes.

2.1 Detection of Anchor Points

Figure 1 shows the significant events in the production of a typical CV unit. Utterances of CV units consists of all or a subset of the following significant speech production events: Closure, burst, aspiration, transition and vowel [11]. The vowel onset point (VOP) is the instant at which the consonant part ends and the vowel part begins in a CV utterance [12]. It is obvious that all the CV units have a distinct VOP in their production [11] [13]. Because every CV utterance has a VOP, the VOPs can be used as anchor points for CV spotting. This approach requires detection of VOPs in continuous speech with a good accuracy. The VOPs of all CV segments in a continuous speech utterance should be detected with minimum deviation. Since labelling will be done only for the segments around the VOPs detected, the effect of any VOP not being detected is that the CV segment around that VOP will not be recognised. Therefore it is important to minimise the number of missing errors by the VOP detection method. The effect of spurious VOPs being detected is that segments around them will also be given to the CV classifier for labelling.

In the method proposed in [13], a multilayer feedforward neural network (MLFFNN) model is trained to detect the VOPs by using the trends in the speech signal parameters at the VOPs. The input layer of the network contains

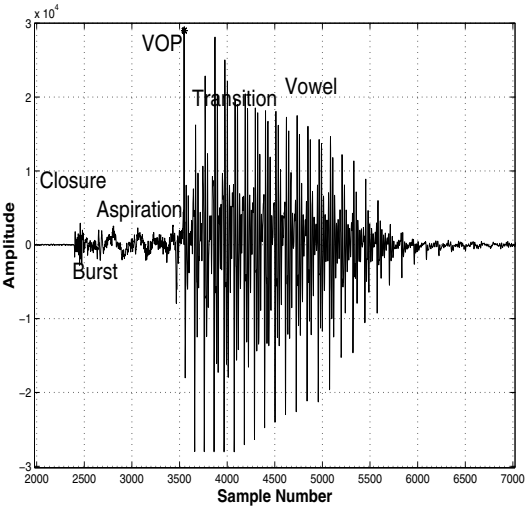

Fig. 1. Significant events in the production of the CV unit /kha/ with VOP at sample number 3549

9 nodes and the output layer has 3 nodes. One of the output nodes is labeled as VOP node to indicate the presence of the VOPs, and the other two nodes are labeled as pre-VOP and post-VOP to indicate the absence of VOPs. The signal energy, residual energy and spectral flatness parameters extracted from two frames around the VOP and the ratio of the parameters in the two frames are used to form an input vector. Two other such vectors are also extracted from each CV utterance. One vector is derived from two frames in the region before the VOP for representing the pre-VOP region. Another vector is derived from two frames in the region after the VOP for representing the post-VOP region. An MLFFNN classifier is trained using the vectors extracted from the three different regions of each utterance. For detection of VOP in a CV utterance using the network trained as above, a parameter vector extracted at every 10 msec is given as input to the network. The parameter vector is extracted from two frames, with one frame starting at the point under consideration and another frame starting 20 msec after this point. Thus the speech signal of a CV utterance is scanned by the network to detect the VOP. The point at which the output for the VOP node of the network is maximum is hypothesised as the VOP of the CV utterance. This method requires a large number of training examples to capture the trends in speech signal parameters at the VOP.

In another method for detection of VOPs, we consider AANN models [3]. A five layer AANN model, shown in Fig. 2, with compression layer in the middle has important properties suitable for distribution capturing, data compression, and extraction of higher order correlation tasks [14] [7]. We explore the distribution capturing of feature vectors by the AANN models to hypothesise the consonant and vowel regions and then detect VOPs in continuous speech. In Section 4, we

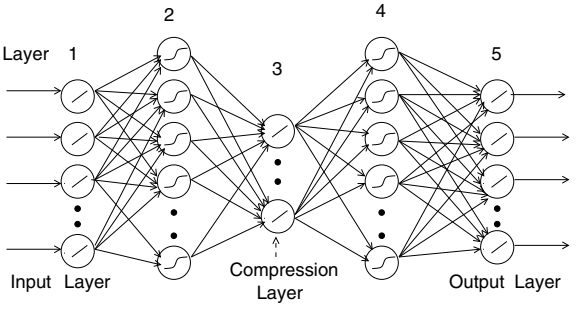

Fig. 2. Five layer AANN model

describe the method used for VOP detection in continuous speech using AANN models.

2.2 Classifier for Recognition of Multilingual CV Segments

Hidden Markov models (HMM) are used in most speech recognition systems. These models use maximum likelihood (ML) approach for training. The incremental model optimization approach in ML framework simplifies the training process, but loses discriminative information in the process [15]. 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 now learn patterns that need to be discriminated. Multilayer feedforward neural network (MLFFNN) models and support vector machine (SVM) models are good at this type of learning since the training involves optimization over entire pattern space [5]. The MLFFNN models have been shown to be suitable for pattern recognition tasks because of their ability to form complex decision surfaces. In order to obtain a better classification performance it is necessary to tune the design parameters such as structure of network, number of epochs, learning rate parameter and momentum. For better generalization, it is necessary to have large amount of training data. But arriving at optimal parameters for complex recognition problem using MLFFNN models is a difficult proportion. SVM models 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. In the next section, we describe the speech corpus used in the studies.

3 Speech Data and Representation

Speech corpus consisting of recording of television broadcast news bulletins for three Indian languages namely, Tamil, Telugu and Hindi is used in our studies.

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

Description	Language			
	Tamil	Telugu	Hindi	Multilingual
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 CVs (Cs:Vs)	324 (27:12)	432 (36:12)	360 (36:10)	432 (36:12)
Number of CV units used for the study	123	138	103	196
Number of CV segments in the training data	44,612	43,491	22,109	1,10,212
Number of CV segments used for training which are covered by number of CV units used for the study	43,541	41,725	20,236	1,05,502
Percentage of number of CV segments used	97.53%	95.93%	91.52%	95.72%
Range of frequency of occurrence for the units in the training data	39 to 1,633	40 to 2,037	40 to 1,264	40 to 2,826
Number of CV segments in the test data covered by number of CV units used for the study	10,293	11,347	4,137	25,777
Speech sentences considered for testing	1,416	1,348	630	3,094

Each bulletin (session) contains 10 to 15 minutes of speech from a single (male or female) speaker. The CV utterances in the corpus are excised and labeled manually. A brief description of the speech corpus used in our studies is given in Table 1. Interpretation of the contents of the table for Tamil language data is as follows: On the whole, 33 bulletins, read by 10 male and 23 female speakers are collected. The data in 27 bulletins read by 8 male speakers and 19 female speakers is used for training. The data in the remaining 6 bulletins by 2 male speakers and 4 female speakers is used for testing. There are 27 consonants and 12 vowels leading to a total of 324 CV units. The CV units have different frequencies of occurrence in the speech corpus. The CV units that occur at least 50 times in the corpus are considered in our studies. This results in a set of 123 CV classes for Tamil language. Out of a total of 44,612 CV segments in the training data, 43,541 segments (*i.e.,* 97.53%) belong to these 123 CV classes. The frequency of occurrence for these classes in the training data varies from 39 to 1,633. The test data includes about 10,293 CV segments belonging to the 123 CV classes. There are 1,416 continuous speech sentences available for testing. A similar description for the speech corpora of Telugu, Hindi and multiple languages is also given in Table 1.

Short-time analysis of the speech signal of the CV utterances is performed using a frame size of 20 msec duration with a shift of 5 msec. Each frame is represented by a parametric vector consisting of 12 mel-frequency cepstral coefficients (MFCC), energy, their first order derivatives and their second order derivatives [16] [17]. The dimension of the parametric vector for each frame is 39.

Models based on SVMs are suitable for classification of fixed dimensional patterns. However, durations of CV utterances vary not only for different classes, but also for a particular CV class. It is necessary to develop a method for representing the CV utterances by fixed dimensional patterns. It is useful to identify the region before the VOP as corresponding to the manner of articulation (MOA), the transition region after VOP to the place of articulation (POA), and the remaining portion to the steady vowel (V). Generally it is difficult to isolate these regions precisely. Moreover, the acoustic characteristics of each region will influence the other regions. Thus, all the three regions need to be represented

together as a single pattern vector [11] [18]. Since the vowel region is prominent in the signal due to its large amplitude characteristics, and also due to its periodic excitation property, it is easy to locate this event compared to other speech production events [13]. The information necessary for classification of CV utterances can be captured by processing a portion of the CV segment containing parts of the closure and vowel region, and all of the burst, aspiration, and transition regions. The closure, burst and aspiration regions are present before the VOP. The transition and vowel regions are present after the VOP. To capture the acoustic characteristics of the CV units, it is necessary to represent each of these units as a sequence of frames, and extract the spectral information corresponding to each frame. A segment of typically 50 to 100 msec duration around the VOP contains most of the information necessary for classification of the CV utterances. This segment can be processed to derive a fixed dimensional pattern, automatically from a varying duration segment of a CV unit [13]. Portions of a CV utterance in the beginning and the end are not included in the fixed duration segment, since they may be affected by the coarticulation effects. From the analysis of broadcast news data it is observed that, the average minimum duration of segments for a CV class is 80 msec. Therefore a 65 msec segment around the VOP is used to represent each CV segment. Once the VOP is detected, five overlapping frames are considered to the left of VOP and five to the right. Thus each CV segment is represented by a 390-dimensional pattern vector. In the next section, we describe the method used for VOP detection in continuous speech using AANN models.

4 System for Detection of VOPs in Continuous Speech Utterances

A five layer AANN model to capture distribution of feature vectors is shown in Fig. 2. In this model the input and output layers have the same number of units, and all these units are linear. For each CV class, two AANN models (one corresponding to the consonant region and the other to the vowel region) are developed. For training the AANN model corresponding to the consonant region, the fifth frame to the left of the manually marked VOP frame is selected from each of the training examples. For training the AANN model corresponding to the vowel region we consider the VOP frame and the fourth frame to the right of VOP frame. The model corresponding to a region of a CV class captures the distribution of feature vectors. The distribution is expected to be different for the consonant and vowel regions of a class. The distribution of feature vectors of a region is captured using a network structure *39L 60N 4N 60N 39L*, where *L* refers to linear units and *N* refers to nonlinear units. The integer value indicates the number of units in that particular layer. The activation function for the nonlinear units is a hyperbolic tangent function. The network is trained using error backpropagation algorithm in pattern mode for 1000 epochs.

For detection of VOPs in continuous speech, each frame is given as input to the pairs of AANN models of all the CV classes. From the evidence available

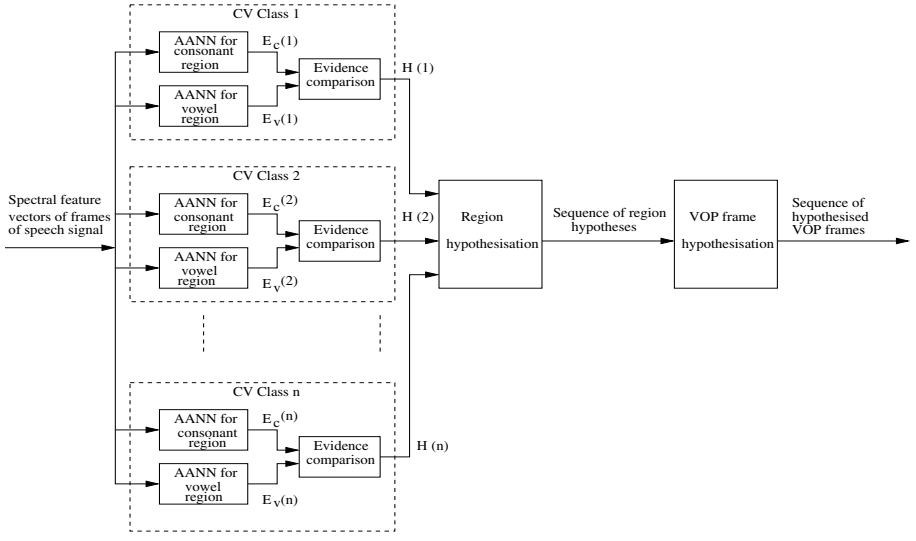

Fig. 3. Block diagram of the system for detection of VOPs in continuous speech. $E_c(k)$ and $E_v(k)$ are the evidence obtained from consonant and vowel region models of k^{th} class respectively. $H(k)$ is hypothesised region of the current frame by the models of class k.

in the outputs of the models of a class, the hypothesised region of the frame is obtained as the region of the model with higher evidence. The hypotheses from the models of different CV classes are used to assign the frame to the consonant or vowel region. In this way we obtain a sequence of region labels for the sequence of frames of the continuous speech utterance. VOP frames are identified as those frames, at which there is a change of labels from consonant to vowel. The block diagram of the system for detection of VOPs in continuous speech utterances is shown in Fig. 3.

We study the performance of the VOP detection method using AANN models. For comparison we consider the method based on MLFFNN model described in Section 2.1 [13]. The performance is measured in terms of the number of matching, missing and spurious hypotheses of VOPs. The VOPs detected with a deviation upto 25 msec are considered as the matching hypotheses. When the deviation of hypothesised VOP is more than 25 msec or there is no hypothesised VOP around an actual VOP, the VOPs of such segments are considered as the missing hypotheses. When there are multiple hypotheses within 25 msec around an actual VOP or the hypothesised VOP does not fall in this range, such hypotheses are considered as spurious ones. For testing we consider the utterances of 120, 120 and 60 sentences selected at random from 1416, 1348, and 630 sentences for Tamil, Telugu and Hindi languages, respectively. These 300 sentences consist of a total number of 3924 syllable-like units corresponding to 1580, 1648 and 696 actual VOPs from sentences of Tamil, Telugu and Hindi languages, respectively. These VOPs have been marked manually. For each utterance the hypothesised VOPs are determined by the MLFFNN and AANN based

Table 2. Comparison of the average performance of different methods for detection of VOPs in continuous speech. The performance is given as a percentage of total number of VOPs in the continuous speech utterances, for the matching, missing and spurious hypotheses.

Method	VOP detection performance (in %)		
	Matching hypotheses	Missing hypotheses	Spurious hypotheses
MLFFNN	68.80	31.19	33.10
AANN	68.62	31.37	6.21

methods. The average performance of different VOP detection methods for the data of three languages is given in Table 2. It is seen from Table 2 that the performance of both the methods is nearly the same for matching case. However, the VOP detection method based on AANN gives significantly less number of spurious VOPs. Many of the missing VOPs in case of AANN based method are observed to be for CV units whose consonants are semivowels, fricatives and nasals.

5 Classification System for Recognition of Multilingual CV Units

In this section, we describe a multilingual system in which data sharing approach is considered for recognition of frequently occurring CV units of three Indian languages. This approach is motivated by the commonality among CV classes across Indian languages. The similar CV classes from different languages are derived from Indian language TRANSliteration (ITRANS) code [19]. The ITRANS code was chosen, as it uses the same symbol across the Indian languages to represent a given sound. A summary of the description of the database used for the development of multilingual CV recognition system is given in the last column of Table 1. The number of CV classes with at least 50 examples in the data set is 123, 138, and 103 for Tamil, Telugu and Hindi respectively, leading to a total of 364 classes. Out of these 364 classes, 27, 25, and 28 classes are unique to Tamil, Telugu and Hindi, respectively. The number of CV classes common to any two languages is 64. There are 52 CV classes common to all the three languages. The union of the set of CV classes in three languages gives a set of 196 CV classes for multilingual data. The number of segments available for training the models of these classes is 1,05,502, and the number of segments in the test data set is 25,777. Thus sharing of data across languages leads to availability of large training data sets, but variability in the data of a class is also increased.

As explained in Section 3, each CV utterance is represented by a pattern vector of dimension 390. To reduce computational complexity, we propose nonlinear compression of the large dimensional input pattern vectors using AANN models [6][7]. The block diagram of the system for recognition of multilingual CV units is shown in Fig. 4. It consists of three stages. In the first stage, the 390-dimensional input pattern vector $\mathbf{x}$ is compressed to a 60-dimensional vector, using an AANN

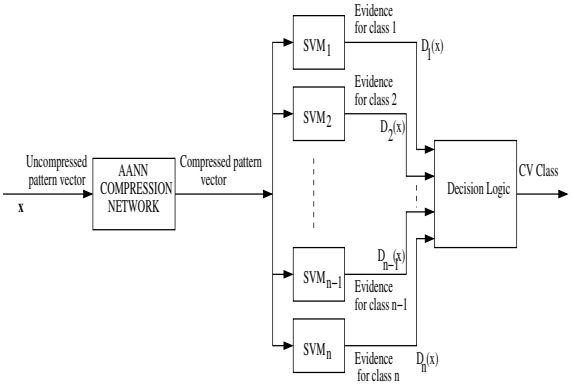

Fig. 4. Block diagram of the multilingual CV recognition system for labelling region around the VOP

with structure *390L 585N 60N 585N 390L*. 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 [20]. SVM models are generated by assigning one model to each class, and training this model by considering data from all the three languages. 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 [5]. The support vectors are those data points that lie closest to the decision surface, and therefore are the most difficult to classify. For a given pattern $\mathbf{x}$ around a VOP, 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 that gives maximum evidence is hypothesised as the class of the pattern $\mathbf{x}$ representing the CV segment around VOP.

The recognition system is developed using the SVM models trained with compressed pattern vectors. The recognition system is also developed using the SVM models trained with 390-dimensional uncompressed vectors. The recognition performance of the SVM models trained with 390-dimensional uncompressed vectors and the models trained with 60-dimensional compressed vectors is given in Table 3. In comparison with uncompressed case, the classification performance is nearly the same for reduced dimension. Thus it is possible to compress the

Table 3. Classification performance of CV recognition systems using compressed and uncompressed pattern vectors in multiple languages

Language	Classification performance (in %)	
	Compressed	Uncompressed
Multilingual	45.31	45.10

Table 4. Comparison of the k-best classification performance for multilingual CV recognition systems

System	$k-$best classification performance (in %)				
	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

390-dimensional pattern vectors to 60-dimensional vectors without affecting the classification performance.

The studies in this section show that it is possible to compress large dimensional pattern vectors to reduced dimensional vectors without affecting the classification performance of the SVM based classifiers. The compression also leads to a significant reduction in the computational complexity of the kernel operations in the SVM models.

The k-best recognition performance of the multilingual system for 196 CV classes is given in Table 4. For comparison, the performance of the hidden Markov model (HMM) based systems is also obtained. A CV segment is analyzed frame by frame, with a frame size of 20 msec and a frame shift of 5 msec. Each frame is represented by a parametric vector consisting of 12 MFCC coefficients, energy, their first order derivatives, and their second order derivatives. In this case, the dimension of each frame is 39. A 5-state, left-to-right, continuous density HMM using multiple mixtures with diagonal covariance matrix is trained for each CV 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 lies between 100 and 500. For other classes, the number of mixtures is 8. All the frames of a CV segment are used in training and testing the HMM based system. The recognition performance of CV segments for the HMM based system is also given in Table 4. It is seen from Table 4 that the SVM based multilingual system performs significantly better than that based on HMMs. The SVMs use discriminative information in the process of learning, whereas HMM models are trained using the maximum likelihood (ML) methods. The ML framework does not use discriminative information. Due to this fact, there is a significant difference (71.03% vs 54.57%) in the 5-best classification performance of SVM and HMM based systems. In the next section, we describe CV recognition system using SVM models for classifying the CV segments around hypothesised VOPs.

6 Spotting CV Units in Continuous Speech

The block diagram of the integrated system for spotting multilingual CV units in continuous speech utterances is given in Fig. 5. The speech signal is given as input to the VOP detection module to locate VOPs in it. The short-time analysis is performed on 65 msec segment around each of the hypothesised VOPs to extract 390-dimensional MFCC based pattern vectors. This pattern vector is compressed using an AANN model. The compressed pattern vector is given to the multilingual CV recognition system to hypothesise the CV class of the current segment. Thus a sequence of hypothesised CV units is obtained for a given speech utterance.

For illustration, we consider a Tamil language continuous speech utterance /kArgil pahudiyilirundu UDuruvalkArarhaL/ consisting of 16 syllables (kAr, gil, pa, hu, di, yi, li, run, du, U, Du, ru, val, kA, rar, haL) whose waveform is shown in Fig. 6(a). The hypothesised region labels obtained using the VOP detection system are shown in Fig. 6(b). The label C corresponds to the consonant region and V to the vowel region. Using the procedure described in Section 4, the VOPs are detected. The hypothesised locations in terms of sample numbers (320, 720, 2440, 3760, 4800, 5560, 6200, 7480, 9480, 11120, 12080, 13240, 14560, 16960) are shown in Fig. 6(c). For comparison we consider manually marked VOP locations (280, 2360, 3800, 4920, 5480, 6320, 7400, **8200**, 9440, 11160, 12080, **12520**, 13200, 14520, **15840**, 16960) shown in Fig. 6(d).

It is seen that there are three VOPs (their sample numbers are indicated in boldface) that have been missed around the locations 8200, 12520, and 15840 corresponding to the syllables /ru/, /ru/, and /ra/, respectively. The VOP at location 720 is hypothesised as spurious VOP. For the segments around the hypothesised VOPs, the five CV class alternatives given by the multilingual CV recognition system (developed in Section 5) are given in Table 5. It is seen that for most of the segments the actual CV class of the segment is present among the alternatives. The correctly identified classes in the CV lattice are written in boldfaces. The segment around the hypothesised location 11120 has been hypothesised as /mu/, where as the actual syllable is /U/. This belongs to the case in which the vowel is in the initial portion of a word. Recognition of only

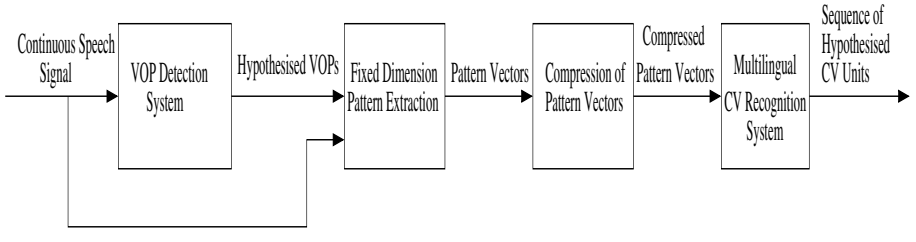

Fig. 5. Block diagram of the multilingual continuous speech recognition system based on spotting CV units

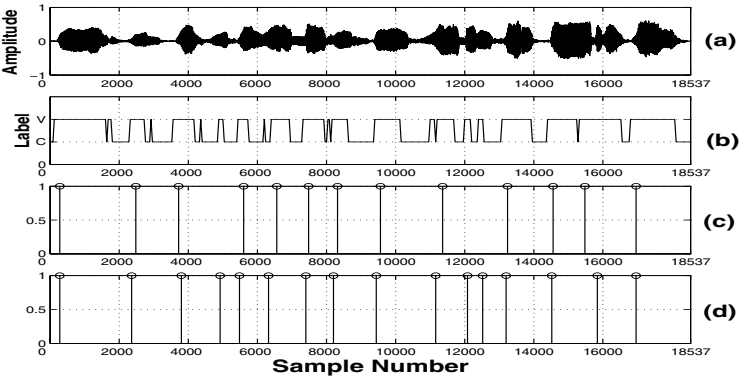

Fig. 6. Plots of the (a) Waveform of the speech signal, (b) Hypothesised region labels for each frame, (c) Hypothesised VOPs, and (d) Manually marked (actual) VOPs for the Tamil language sentence /kArgil pahudiyilirundu UDuruvalkArarhaL/

Table 5. The classes hypothesised by the multilingual CV classifier for a continuous speech utterance /kArgil pahudiyilirundu UDuruvalkArarhaL/. The alternative classes for the segment around a hypothesised VOP are given in a row of the lattice. The entries in the last column represent position of actual CV in hypothesised alternatives.

VOP locations (in sample numbers)		Lattice of hypothesised CVs					Actual	Position
Actual	Hypothesised	1	2	3	4	5	syllable	
280	320	pA	**kA**	vA	ha	shu	kAr	2
—	720	kA	pA	hA	na	pa	—	—
2360	2440	**gi**	yE	hi	ya	yai	gil	1
3800	3760	hA	pA	**pa**	sA	sa	pa	3
4920	4800	**hu**	gu	mu	vu	pu	hu	1
5480	5560	bI	vi	Ti	Ni	**dI**	di	5
6320	6200	**yi**	lA	li	zi	tI	yi	1
7400	7480	**li**	ni	ru	ja	lai	li	1
8200	—	VOP Missed					run	—
9440	9480	**du**	Ru	ja	dE	rA	du	1
11160	11120	**mu**	kU	va	pO	vA	U	1
12080	12080	**Du**	da	dA	nA	tu	Du	1
12520	—	VOP Missed					ru	—
13200	13240	**va**	da	kai	hi	vA	val	1
14520	14560	**kA**	ka	ga	cha	zA	kA	1
15840	—	VOP Missed					rar	—
16960	16960	**ha**	kA	ka	ga	sa	haL	1

vowels is not addressed in the current studies. All the classes hypothesised by the recognition system are of type CV.

We study the performance of the spotting approach for recognition of CV units for a large number of sentences in three Indian languages. For testing we consider 300 sentences from different languages consisting of a total number of

3924 syllable-like units corresponding to 1580, 1648 and 696 actual VOPs from sentences of Tamil, Telugu and Hindi languages, respectively. These VOPs have been marked manually. For each sentence the hypothesised VOPs are determined by the AANN method explained in Section 4. The VOPs that are detected with a deviation upto 25 msec are about 68.62% and there are about 6.21% of spurious VOPs. About 74.63% of the CV segments have been correctly recognised in five alternatives by spotting the CV segments around the detected VOPs.

7 Summary and Conclusions

In this paper, we have addressed the issues in spotting based approach for recognition of consonant-vowel (CV) units in multiple languages. The approach is based on using the vowel onset points (VOPs) as anchor points and then classifying the segments around VOPs using a classifier. Autoassociative neural network (AANN) models are used for detecting VOPs in continuous speech. The methods for minimising the number of missing VOPs have to be explored. We use support vector machine (SVM) based classifier for recognition of CV segments around the hypothesised VOPs. To reduce the computational complexity of kernel operations in the SVM models, we perform nonlinear compression using AANN models for compression of pattern vectors. The results show that it is possible to compress the 390-dimensional pattern vectors to 60-dimensional vectors without affecting the classification performance. We proposed a data sharing approach for the development of multilingual CV recognition system. Though the variability among the data of a class is more and the number of CV classes is larger for the multilingual system, it has less effect on the recognition performance when SVMs are used for classification. However, classification performance of the hidden Markov model (HMM) based system is affected more by the large number of classes. The hypothesised CV sequence can be processed to perform word-level matching and sentence-level matching to recognise complete sentences.

References

1. L. R. Rabiner and B. -H. Juang: Fundamentals of Speech Recognition. PTR Prentice Hall, Englewood Cliffs, New Jersey (1993)
2. 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. (1987) 369–372
3. S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana: Detection of vowel onset points in continuous speech using autoassociative neural network models. In: Proc. Eighth Int. Conf. Spoken Language Processing (INTERSPEECH 2004 - ICSLP). (2004) 1081–1084
4. S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana: Acoustic model combination for recognition of speech in multiple languages using support vector machines. In: Proc. IEEE Int. Joint Conf. Neural Networks (Budapest, Hungary). Volume 4(4). (2004) 3065–3069

5. S. Haykin: Neural Networks: A Comprehensive Foundation. Prentice-Hall International, New Jersey (1999)
6. 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). (2003) 156–159
7. K. I. Diamantaras and S. Y. Kung: Principal Component Neural Networks, Theory and Applications. John Wiley and Sons, Inc., New York (1996)
8. S. Roukos, R. Rohlicek, W. Russel, and H. Gish: Continuous hidden Markov modelling for speaker-independent word spotting. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing. (1989) 627–630
9. C. Chandra Sekhar and B. Yegnanarayana: Neural network models for spotting stop consonant-vowel (SCV) segments in continuous speech. In: Proc. Int. Conf. Neural Networks. (1996) 2003–2008
10. S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana: Spotting consonant-vowel units in continuous speech using autoassociative neural networks and support vector machines. In: Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (Sao Luis, Brazil). (2004) 401–410
11. C. Chandra Sekhar: Neural Network Models for Recognition of Stop Consonant-Vowel (SCV) Segments in Continuous Speech. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras (1996)
12. S. V. Gangashetty and S. R. Mahadeva Prasanna: Significance of vowel onset point for speech recognition using neural network models. In: Proc. Fifth Int. Conf. Cognitive and Neural Systems (Boston, USA). (2001) 24
13. J. Y. Siva Rama Krishna Rao, C. Chandra Sekhar, and B. Yegnanarayana: Neural networks based approach for detection of vowel onset points. In: Proc. Int. Conf. Advances in Pattern Recognition and Digital Techniques, Calcutta. (1999) 316–320
14. B. Yegnanarayana and S. P. Kishore: AANN-An alternative to GMM for pattern recognition. Neural Networks **15** (2002) 459–469
15. H. Bourlard and N. Morgan: Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, Boston (1994)
16. 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 **28** (1980) 357–366
17. S. Furui: On the role of spectral transition for speech perception. J. Acoust. Soc. Am. **80(4)** (1986) 1016–1025
18. C. Chandra Sekhar and B. Yegnanarayana: A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances. IEEE Trans. Speech and Audio Processing **10** (2002) 472–480
19. A. Chopde: ITRANS Indian Language Transliteration Package Version 5.2. (Source, http://www.aczone.com/itrans/)
20. 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). (2002) 171–185