2004 IEEE Workshop on Machine Learning for Signal Processing

SPOTTING CONSONANT-VOWEL UNITS IN CONTINUOUS SPEECH USING AUTOASSOCIATIVE NEURAL NETWORKS AND 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 propose an approach for continuous speech recognition by spotting consonant-vowel (CV) units. The main issues in spotting CV units are the location of anchor points and labelling 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 consider support vector machine (SVM) based classifiers due to their ability of generalisation from limited training data and also due to their inherent discriminative learning. The CV spotting approach for continuous speech recognition has been demonstrated for sentences in 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 [7]. 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 models as compositions of subword

0-7803-8608-6/04/$20.00 ©2004 IEEE

unit models, and recognising sentences by performing word-level matching and sentence level matching using word models and language models respectively [7]. 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 continuous speech recognition by spotting subword units. Specifically, we develop 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 distribution capturing ability of autoassociative neural network (AANN) models is explored for detection of VOPs in continuous speech [14]. 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 [10]. 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 [8]. 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. We demonstrate the CV spotting based approach to continuous speech recognition for sentences in Indian languages.

The paper is organised as follows: In Section 2, we discuss the issues in spotting CV units. The system for spotting CV units in continuous speech is described in Section 3. In Section 4, 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 is also presented in this section.

2. ISSUES IN SPOTTING 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 [11]. The main limitation of this strategy based on scanning is that a large number of spurious hypotheses are given by the spotting system [2]. For spotting CV units in continuous speech, we consider an approach based on detection of VOPs and labelling the segments around the VOPs using a CV classifier. The main issues in spotting CV units in the proposed approach are location of anchor points and labelling the regions around these anchor

points using suitable classifiers.

2.1 Location of anchor points

Utterances of CV units consist of all or a subset of the following significant speech production events: closure, burst, aspiration, transition and vowel. The vowel onset point (VOP) is the instant at which the consonant part ends and the vowel part begins in a CV utterance. Since the vowel region is prominent in the signal due to its large amplitude characteristics and periodic excitation property, it is easy to locate this event compared to other speech production events. 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. 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 [5], 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. We consider autoassociative neural network (AANN) models for detection of VOPs. In an AANN model the input and output layers have the same number of units, and all these units are linear. A five layer AANN model, with compression layer in the middle has important properties suitable for distribution capturing, data compression, and extraction of higher order correlation tasks [1] [6]. 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 3.1, we describe the method used for VOP detection in continuous speech using AANN models.

2.2 Classifier for recognition of 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 [4]. 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 models and support vector machine (SVM) models are good at this type of learning since the training involves optimization over entire pattern space [10]. 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.

SVM models have attained prominence due to their inherent discriminative learning and generalization capabilities from the limited training data [13]. 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. SVM models are suitable for handling patterns of fixed dimension. For this purpose, a segment of fixed duration around the VOP that contains most of the information necessary for classification of CV utterances can be processed to derive a fixed dimension pattern. 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. In the next section, we describe CV recognition system using SVM models for classifying the CV segments around the hypothesised VOPs.

3. SYSTEM FOR SPOTTING CV UNITS

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 for these three languages is given in Table 1. Each bulletin contains 10 to 15 minutes of speech from a single (male or female) speaker. The CV utterances in the database are segmented and labeled manually. The CV units have different frequencies of occurrence in the database. We consider a set of CV classes that have a frequency of occurrence greater than 50. Short-time analysis of the speech signal of the CV utterances is performed using frames 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 [9]. Thus the dimension of each frame is 39. Systems are developed for spotting CV units for the data of each of the three languages.

3.1 System for detection of VOPs

A five layer AANN model to capture distribution of feature vectors is shown in Fig. 1. For each CV class, two AANN models are developed. For train-

TABLE 1: DESCRIPTION OF BROADCAST NEWS SPEECH CORPUS USED IN STUDIES.

Description	Language		
	Tamil	Telugu	Hindi
Number of bulletins	33	20	19
Number of bulletins used for training	27	16	16
Number of bulletins used for testing	6	4	3
CV classes used for the study	123	138	103
CV segments used for training	43,541	41,725	20,236
Speech sentences considered for testing	1,416	1,348	630

ing 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 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. 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.

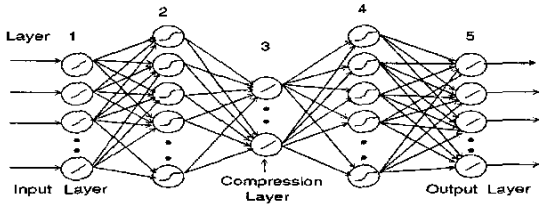

Figure 1: Five layer AANN model.

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 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. 2.

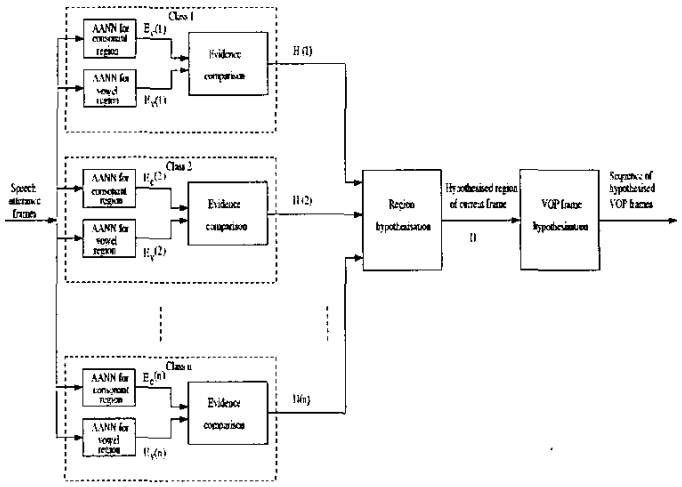

Figure 2: Block diagram of the system for detection of VOPs in continuous speech. $E_c(k)$ and $E_v(k)$ are the evidence obtained from consonant and vowel region models of k^{th} class respectively. $H(k)$ is hypothesised region of the current frame by the models of class k. H is the hypothesis of the current frame.

3.2 Classification system for recognition of CV units

For fixed dimension representation of each CV utterance of the training data, we consider 65 msec around the VOP. Five overlapping frames are considered to the left of VOP and five to the right of VOP, with a shift of 5 msec. Thus, the pattern vector for each CV utterance is a 390-dimension vector formed by concatenating the feature vectors of 10 successive frames. To reduce computations complexity, we propose nonlinear compression of the large dimension input pattern vectors using AANN models [12][6]. The block diagram of the system for recognition of CV units is shown in Fig. 3. 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*. 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 [3]. 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 [10]. The support vectors are those data points that lie closest to the decision surface, and therefore the most difficult to classify. For a given pattern $\mathbf{x}$ around the 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.

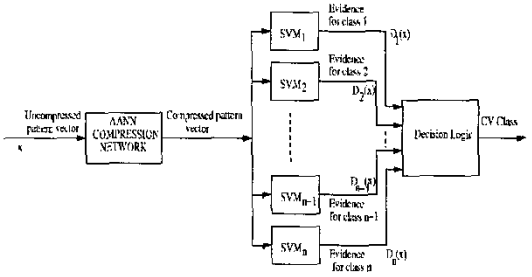

Figure 3: Block diagram of the CV recognition system using AANN model for compression of pattern vectors followed by SVM models for classification.

The block diagram of the integrated system is given in Fig 4. 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-dimension MFCC based pattern vectors. This pattern vector is compressed to 60-dimension using AANN compression network. The compressed pattern vector is given to the CV recognition system to hypothesise the CV class of the current segment. Thus a sequence of hypothesised CV units is obtained for the given speech utterance.

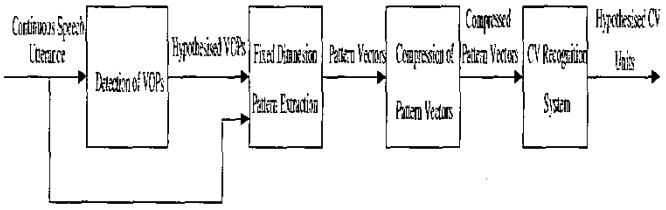

Figure 4: Block diagram of the continuous speech recognition system based on spotting CV units.

4. SPOTTING CV UNITS IN CONTINUOUS SPEECH

For illustration, we consider a 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. 5 (a). The hypothesised region labels using AANN models are shown in Fig. 5 (b). The label C corresponds to consonant region and V to vowel region. Using the procedure described in Section 3.1, the VOPs are detected. The hypothesised locations in terms of sample numbers (280, 2480, 3720, 5600, 6560, 7480, 8320, 9560, 11360, 13240, 14560, 15480, 16960) are shown in Fig. 5 (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. 5 (d).

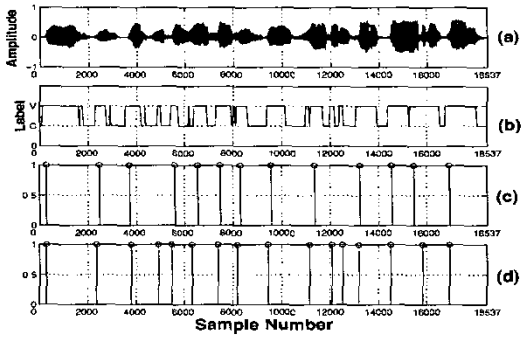

Figure 5: Plots of the (a) Waveform of the speech signal, (b) Hypothesised region labels for each frame, (c) Hypothesised VOPs, and (d) Manually marked (actual) VOPs for the Tamil language sentence /kArgil pahudiyilirundu UDuruvalkArarhaL/.

It is seen that there are four VOPs (their sample numbers indicated in boldface) that have been missed around the locations 4920, 12080, 12520, and 15840 corresponding to the syllables /**hu**/, /Du/, /ru/, and /ra/ respectively. It is seen that there are fewer than 8 vowel region hypotheses around these locations. The VOP at location 15480 is hypothesised as spurious VOP. For hypothesised VOPs, the CV alternatives hypothesised by the recognition system are given in Table 2. For most of the hypothesised VOPs, the actual CV class of the segment around the VOP is present in the alternatives. The correctly identified classes in the CV lattice are written in boldfaces. The segment around the location 11360 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 word. Recognition of only 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

TABLE 2: THE CLASSES HYPOTHESISED BY THE 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. THE ENTRIES IN THE LAST COLUMN REPRESENT POSITION OF ACTUAL CV IN HYPOTHESISED ALTERNATIVES.

VOP locations		Lattice of hypothesised CVs					Actual	Position
Actual	Hypothesised	1	2	3	4	5	syllable	
280	280	**kA**	hA	pA	ka	kai	kAr	1
2360	2480	**gi**	yi	yE	ya	hi	gil	1
3800	3720	hA	**pa**	pA	sa	kA	pa	2
4290	—	VOP Missed					hu	-
5480	5600	**dI**	bI	TI	NI	vI	dI	1
6320	6560	tI	yE	yai	**yi**	kA	yi	4
7400	7480	ni	**li**	ru	la	ja	li	2
8200	8320	**ru**	Ru	ra	Ra	NA	run	1
9440	9560	Ru	ru	**du**	ha	NA	du	3
11160	11360	**mu**	pu	ha	mA	ku	U	1
12080	—	VOP Missed					Du	-
12520	—	VOP Missed					ru	-
13200	13240	**va**	vai	da	kai	hi	val	1
14520	14560	**kA**	ka	ga	pa	zA	kA	1
—	15480	pa	ta	di	ka	ha	—	-
15840	—	VOP Missed					rar	-
16960	16960	**ha**	sa	ka	TA	LA	haL	1

units for a large number of sentences in three Indian languages. For testing we consider 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 sentence the hypothesised VOPs are determined by the method explained in Section 3.1. The VOPs that are detected with a deviation upto 25 msec are about 68.19% and there are about 6.65% of spurious VOPs [14]. About 65% of the CV segments have been correctly recognised in five alternatives by spotting the CV segments around the detected VOPs.

5. SUMMARY AND CONCLUSIONS

In this paper, we have addressed the issues in consonant-vowel (CV) spotting based approach for continuous speech recognition. 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 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. The hypothesised CV sequence can be processed to perform word-level matching and sentence-level matching to recognise complete sentences.

REFERENCES

[1] B. Yegnanarayana and S. P. Kishore, "AANN-An alternative to GMM for pattern recognition," **Neural Networks**, vol. 15, no. 3, pp. 459–469, Apr. 2002.

[2] 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, pp. 2003–2008.

[3] 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.

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

[5] 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**, Dec. 1999, pp. 316–320.

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

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

[8] 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.

[9] 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.

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

[11] 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, pp. 627–630.

[12] 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.

[13] S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Acoustic model combination for recognition of speech in multiple languages using support vector machines," in **to be presented in IEEE Int. Joint Conf. Neural Networks (Budapest, Hungary)**, Jul. 2004.

[14] S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Detection of vowel onset points in continuous speech using autoassociative neural network models," in **to be presented in Eighth Int. Conf. Spoken Language Processing (INTERSPEECH 2004-ICSLP)**, Oct. 2004.