Extraction of Fixed Dimension Patterns from Varying Duration Segments of Consonant-Vowel Utterances

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

Classification models based on multilayer perceptron (MLP) or support vector machine (SVM) have been commonly used for complex pattern classification tasks. These models are suitable for classification of fixed dimension patterns. However, durations of consonant-vowel (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 patterns of fixed dimension. For CV utterances, vowel onset point (VOP) is the instant at which the consonant part ends and the vowel part begins. Important information necessary for classification of CV utterances is present in the region around the VOP. A segment of fixed duration around the VOP can be processed to extract a pattern of fixed dimension to represent a CV utterance. Accurate detection of vowel onset points is important for recognition of CV utterances of speech. In this paper, we propose an approach for detection of VOP, based on dynamic time alignment between a reference pattern of a CV class and the pattern of an utterance of that class. The results of studies show that the hypothesised VOPs using the proposed approach have less deviation from their actual locations.

1. INTRODUCTION

In this paper, we address the issues in the detection of vowel onset points (VOPs) in syllable-like units. Speech recognition involves transforming the input speech into a sequence of units called symbols, and converting the symbol sequence into a text corresponding to the message conveyed by the speech signal. Syllable-like subword units such as consonant-vowel (CV) units are important from the point of view of speech production and perception [1].

The number of CV classes in a language is large (more than 300), and many of these classes have similar acoustic features. Classification models based on multilayer perceptron (MLP) or support vector machines (SVM) have been commonly used for complex pattern classification tasks. The CV utterances, by nature of their production, have varying durations. However the classification models based on MLP or SVM are capable of handling only patterns of fixed dimension. Therefore, it is necessary to derive the fixed dimension patterns from CV utterances.

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 also due to its 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. A segment of fixed duration (50 to 100 msec) around the VOP contains most of the information necessary for classification of CV utterances. This segment can be processed to derive a fixed dimension pattern automatically. 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 this approach of pattern representation it is necessary to develop a method for detection of VOPs in CV utterances with a good accuracy. In this paper, we address the issues in the detection of VOPs in CV utterances.

The paper is organized as follows: In Section 2, we briefly review the VOP detection methods. The first method is based on strengths of instants of significant excitation. The second method is based on a neural network model,which captures the trend in the signal parameters before and after the VOP. Then we propose a method for detection of VOP using dynamic time alignment. In Section 3, we study the performance of these three VOP detection methods in terms of average deviation of hypothesised VOPs from their actual locations. The effect of deviation in VOP detection on recognition of CV utterances is studied in Section 4. In this section, we also discuss the effect of deviation in VOP detection on the complexity of SVM classifiers.

0-7803-8243-9/04/$17.00 © 2004 IEEE

2. METHODS FOR DETECTION OF VOPS

The method proposed in [2] detects VOPs by identifying the points at which there is a rapid increase in the vowel strength. The vowel strength is computed using the difference in the energy of each of the peaks in the amplitude spectrum and the energy of a dip associated with the peak. Speech segments with the duration of pitch periods are analysed to obtain the amplitude spectrum and compute the vowel strength. This method requires unvoiced/voiced classification of the speech signal. The methods proposed in [3] and [4] first classify the speech signal into voiced/unvoiced/silence regions using a neural network classifier, and then label the voiced regions as vowel and nonvowel regions. Segmentation of continuous speech into vowel-like and nonvowel-like regions was proposed in [5]. Features such as energy, ratio of the high frequency energy to the low frequency energy, ratio of the volumes of back and total cavities of vocal tract are used. All these methods first classify the segments of speech as vowel or nonvowel regions, and then detect the VOP by identifying the point at which the vowel region begins.

2.1 Approach based on instants of significant excitation

The method proposed in [6] computes the strengths of instants of significant excitation for a given CV utterance. Next, the instant at which there is a significant change in the strengths is detected. This instant is hypothesised as the VOP. These steps are implemented as follows: Initially, the signal is preemphasised. LP analysis of order 10 is performed on frames of 20 msec duration, with a frame shift of 10 msec. LP residual is obtained by inverse filtering the speech signal. The Hilbert envelope of the LP residual is computed. The strengths of excitation are obtained by convolving the Hilbert envelope with a Gabor filter, and the location of the maximum is hypothesised as VOP.

2.2 Neural network based approach

In the method proposed in [7], an MLP 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 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 MLP 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. Now, we discuss our proposed method for detection of VOP using dynamic time alignment. This method requires a single reference pattern per class.

2.3 Dynamic time warping based approach

Dynamic programming is used in speech processing applications for time alignment and normalization to compensate for variability in speaking rate in template-based systems [8]. Let us consider two speech patterns X and Y, representing the spectral sequences $(\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_i}, \cdots, \mathbf{x_M}$)and $(\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_j}, \cdots, \mathbf{y_N}$, where $\mathbf{x_i}$ and $\mathbf{y_j}$ are parameter vectors of short-time acoustic features. Dissimilarity between the two sequences X and Y for a particular path ϕ is given by $d_\phi(X, Y)$. Time normalization of X and Y is obtained by finding the best temporal match given by the minimum dissimilarity $d(X, Y)$, defined as:

$$d(X,Y) \equiv \min_{\phi} \quad (d_\phi(X,Y)) \qquad (1)$$

When X and Y represent CV utterances of the same class, the choice of the best path implies that the dissimilarity is measured based on the best possible alignment between the different regions of the two utterances. In dynamic time alignment, a set of local continuity constraints are imposed on the warping function, which does not result in the omission of any important information bearing events of the CV utterance. The definition of the dissimilarity measure given by (1) involves a minimization process that can be effectively solved by dynamic programming. Hence the name dynamic time warping (DTW). While comparing CV utterances of the same class, the optimal warping path that provides the best match is expected to lead to the alignment of VOP frames.

Following are the steps used for detection of VOP in our proposed approach:

1. A reference (template) pattern is selected for each CV class and the VOP frame in it is identified manually.
2. At the beginning and end of each CV utterance, 20 msec silence is inserted. This helps in satisfying the endpoint constraints.

3. Each CV utterance is processed to obtain a sequence of frames using the frame size of 10 msec and the frame shift of 1 msec. Each frame is represented by 5 weighted linear prediction cepstral coefficients (WLPCCs) obtained using a 3^{rd} order LP analysis. These specifications help in better resolution (smooth warping path) and speaker independence.
4. Optimal warping path between the reference pattern and the given pattern of a class is obtained using DTW algorithm.
5. The frame of the given pattern that aligns with the VOP frame in the reference pattern is hypothesised as the VOP frame of the given pattern.

Fig. 1 shows a typical warping path that is a result of dynamic time alignment between a reference pattern of class /Ra/, and an input pattern of the same class. For the reference pattern the VOP is identified manually at sample number 637 (frame index $i_x = 80$). From the warping path, the VOP of the input pattern is hypothesised at sample number 665 (frame index $i_y = 84$). For this case, Fig. 2 indicates the hypothesised VOPs using the above three methods. The VOPs hypothesised by the instants based approach and the neural network (NN) based approach are at sample numbers 782 and 760 respectively. In the next section, the performance of the above three methods of VOP detection is compared in terms of average deviation of hypothesised VOPs from their actual locations.

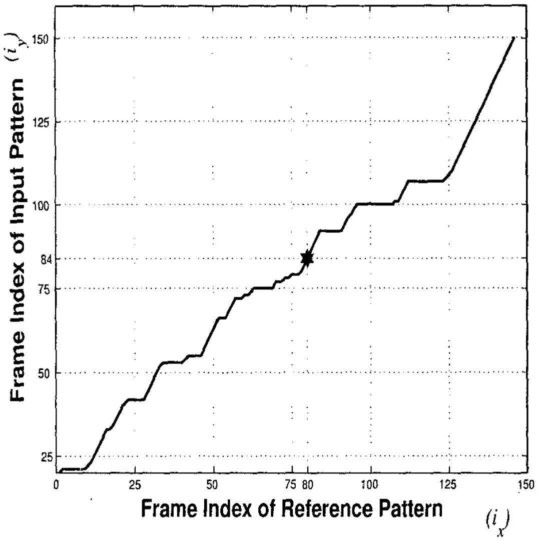

Figure 1. Warping path between the reference pattern and an input pattern for class /Ra/. The VOP frame index for the reference pattern is $i_x = 80$. The corresponding frame index of the input pattern is $i_y = 84$, which is marked as $\star$ on the warping path. This frame is hypothesised as VOP frame for the input utterance.

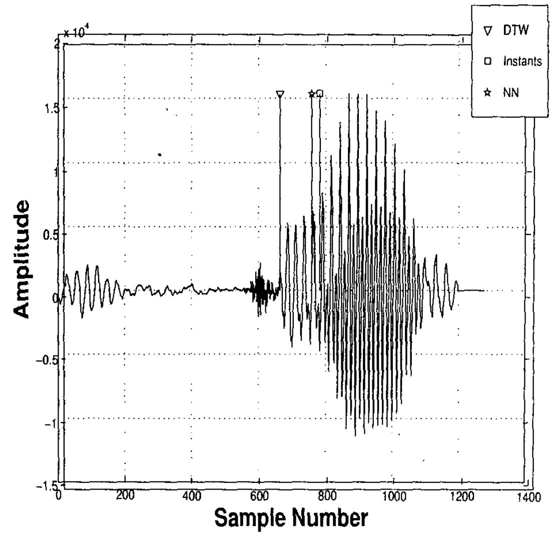

Figure 2. Hypothesised VOPs for a CV utterance of class /Ra/ using different methods. The VOPs hypothesised by the instants based approach, neural network (NN) based approach, and DTW based approach are at sample numbers 782, 760, and 665 respectively. The actual (manually marked) VOP is at sample number 666.

3. STUDIES ON DETECTION OF VOPS

In this section we study the performance of the above three VOP detection methods in terms of average deviation of hypothesised VOPs of CV utterances from their actual (manually marked) locations. A speech database consisting of recordings of TV news bulletins in Tamil language is used in our studies. A summary of the database for this language 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 varying frequencies of occurrence in the database. We consider a set of 123 CV classes that occur more than 50 times in the database.

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

Number of bulletins	33
Number of bulletins used for training	27
Number of bulletins used for testing	6
Number of CV classes used for the study	123
Number of CV segments used for training	43,541
Range of frequency of occurrence for the classes in the training data	39 to 1,633
Number of CV segments used for testing	10,293

For each CV class, the actual VOP has been manually marked for 10 randomly chosen utterances, to study the performance of the VOP detection methods. For each CV utterance, the VOP hypothesised by a method is determined as explained in the previous section. The average deviation $D(k)$ of the hypothesised VOPs from their actual locations for a set of 10 utterances of a CV class k is computed as follows:

$$D(k) = \frac{1}{10}\sum_{i=1}^{10}(A(k,i) - H(k,i)), \quad k = 1,2,\ldots,123 \quad (2)$$

where $A(k,i)$ is the actual VOP and $H(k,i)$ is the hypothesised VOP by a method, for i^{th} randomly selected CV utterance of class k.

The average deviation of hypothesised VOPs from their actual VOPs for different classes is plotted in Fig. 3. It is seen from Table 2 that, the average deviation is in the range of ± 25 msec for a large number of classes when the VOP is hypothesised using the DTW based approach. Since this approach is guided by reference pattern of a class, VOP detection is more accurate than other two approaches in many cases. In the next section, we study the effect of deviation in VOP detection on recognition performance of CV utterances and complexity of the classifier.

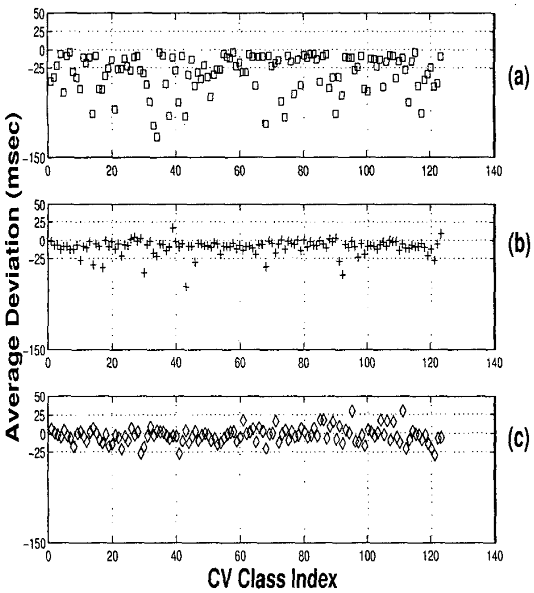

Figure 3. Average deviation of hypothesised VOPs from actual VOPs using (a) Instants of excitation. (b) Neural network. (c) DTW based VOP detection methods.

Table 2. Extent of average deviation (msec) for different VOP detection methods.

VOP detection method	Number of classes having average deviation	
	<=± 25 msec	>± 25 msec
Instants of excitation	61	62
Neural network	113	10
DTW	118	5

4. EFFECT OF DEVIATION IN VOP DETECTION ON RECOGNITION OF CV UNITS

The main purpose of detection of VOPs in CV utterances is to use the VOP as an anchor point for extraction of fixed dimension patterns from varying duration utterances. In this section we study the effect of deviation in hypothesisd VOP on recognition of CV units. A summary of details of the database used in the studies is given in Table 1. We have considered a set of 123 CV classes for which the training data includes 43,541 CV segments and the test data includes 10,293 CV segments.

In our studies on the recognition of CV units, we consider support vector machine (SVM) models for classification [9]. The performance of the above three methods of VOP detection is compared in terms of recognition performance of CV utterances, and complexity of the classifier. For each VOP detection method, we conducted four sets of studies corresponding to four different types of CV representation around the hypothesised VOP. The four different types of pattern representation are described below:

- Type-1 representation: In this type of pattern representation, one frame of 20 msec is considered to the left of VOP, and four overlapping frames, each of 20 msec, are considered to the right of VOP, with a shift of 5 msec. Thus a 55 msec segment anchored around the VOP hypothesised by the corresponding VOP detection method is used to represent each CV utterance. Each frame is represented by 10 WLPCCs which are obtained using an 8^{th} order LP analysis [8]. Thus the pattern vector for each CV utterance is a 50-dimension vector formed by concatenating the five 10-dimension WLPCCs.

- Type-2 representation: Five overlapping frames are considered to the left of VOP, and five to the right of VOP, with a shift of 5 msec. The duration of each frame is 20 msec. Thus a 65 msec segment around the VOP is used to represent each CV utterance. Each frame is represented by 10 WLPCCs which are obtained using an 8^{th} order LP analysis. The pattern vector for each CV utterance is a 100-dimension vector formed by concatenating the ten 10-dimension WLPCCs.

- Type-3 representation: For each of the frames extracted for Type-2 representation, the first order derivatives (delta coefficients) and the second order derivatives (acceleration coefficients) of the WLPCCs are appended to get ten frames each of dimension 30. Thus, in this type of representation each CV utterance is of 300-dimension.

- Type-4 representation: In this type of representation, each frame is represented by a parametric vector consisting of 12 mel-frequency cepstral coefficients, energy, their first order derivatives and their second order derivatives. Thus the dimension of each frame is 39. We consider ten frames as explained in Type-2 representation. Thus each CV utterance is represented by a 390-dimension vector.

Recognition performance of CV units for different VOP detection methods is given in Table 3. The table gives the percentage of correct classification of test utterances. The results show that the VOP detection by DTW performs slightly better than the other two methods for all four types of pattern representation. When the hypothesised VOP has a lesser deviation from the actual VOP, the pattern representation covers consonant, transition, and vowel regions. When the hypothesised VOP is to the right side (negative deviation) of the actual VOP with a greater deviation, pattern representation may lead to loss of information from the consonant and transition regions. The formant transitions in the transition region ,provide the information about the consonant. When this region is omitted, only the vowel region may be represented. This leads to confusion among the patterns resulting in poorer recognition. It is seen from Fig. 3 that, there are large number of classes having positive deviation for the case of DTW based method, large number of classes are having less negative deviation for neural network based method, and large number of classes are having more negative deviation for instants based method. This is reflected in the recognition performance of CV units, as shown in Table 3.

Table 3. Comparison of the classification performance for different VOP detection methods. SVM models are used for classification of CV units.

VOP detection method	CV pattern representation			
	Type-1	Type-2	Type-3	Type-4
Instants of excitation	23.42	31.49	37.15	43.17
Neural network	27.90	35.18	41.82	46.58
DTW	32.08	36.90	43.06	48.73

The complexity of SVM models in terms of average number of support vectors per class for different VOP detection methods is given in Table 4. Support vectors are those training patterns that lie closest to the margin of separation of two classes and are therefore most difficult to classify. They are the more confusable patterns. The results show that the complexity of the SVM model is always less for the case of DTW in comparison with the other two methods for all four types of pattern representation. Using Type-4 CV pattern representation, the number of support vectors used by SVM models for each class for different VOP detection methods is given in Fig. 4. From this figure, it is seen that the number of support vectors used by SVM models for any class is lesser for the case of VOPs hypothesised by DTW approach. More number of patterns fall near to the margin of separation when they are extracted around the hypothesised VOPs by other two methods. Hence they are more confusable. From the complexity of the SVM models for different VOP detection methods, it is seen that the patterns extracted around the VOPs hypothesised by DTW approach represent CV utterances better and hence cause lesser confusion during classification.

Table 4. Comparison of the complexity of the SVM classifiers for different VOP detection methods. The complexity is given as the average number of support vectors per class.

VOP detection method	CV pattern representation			
	Type-1	Type-2	Type-3	Type-4
Instants of excitation	1459	1214	1137	3330
Neural network	1377	1130	1070	3217
DTW	1282	1058	1022	1378

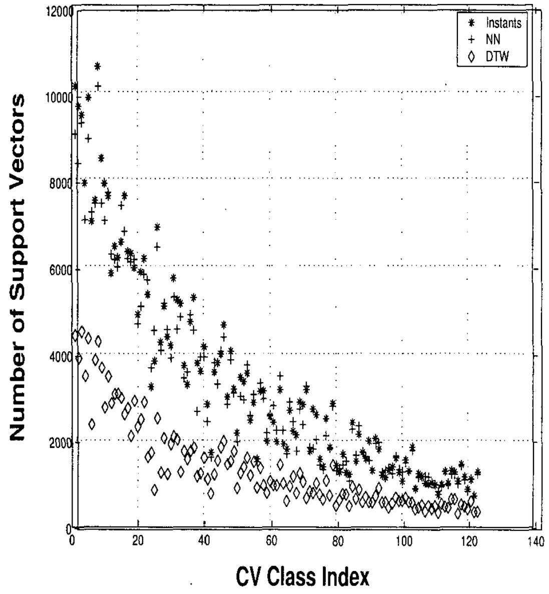

Figure 4. Number of support vectors in SVM models for different VOP detection methods.

5. SUMMARY AND CONCLUSIONS

In this paper we have proposed a method for detection of vowel onset points (VOPs) for consonant-vowel (CV) utterances using dynamic time warping (DTW). The VOPs are used as an anchor points for extraction of fixed dimension patterns from varying duration segments of CV utterances. It has been demonstrated that the DTW based method gives a slightly better performance. Since our approach is guided by a reference pattern of a class, the hypothesised VOP has lesser deviation from its actual location in many utterances. Selection of reference patterns for classes plays an important role for the proposed method. It is necessary to develop the suitable methods for detection of VOPs in utterances with multiple CV segments so that the corresponding subword units can be spotted in continuous speech.

REFERENCES

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

[2] D. J. Hermes, "Vowel-onset detection," J. Acoust. Soc. Am., vol. 87, no. 2, pp. 866–873, Feb. 1990.

[3] A. Bendiksen and K. Steiglitz, "Neural networks for voiced/unvoiced speech classification," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, May. 1990, pp. 521–524.

[4] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifiers," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 1, pp. 250–255, Apr. 1993.

[5] H. Kasuya and H. Wakita, "An approach to segmenting speech into vowel and nonvowel-like intervals," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 27, no. 1, pp. 319–327, Aug. 1979.

[6] S. R. Mahadeva Prasanna and B. Yegnanarayana, "Detection of vowel onset point using source features," Communicated to Speech Communication, 2002.

[7] 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.

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

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