Neural Networks based Approach for Detection of Vowel Onset Points

J.Y.Siva Rama Krishna Rao, C.Chandra Sekhar and B.Yegnanarayana
Department of Computer Science and Engineering
Indian Institute of Technology, Madras - 600 036, India
Email: chandra@iitm.ernet.in

Abstract

Detection of Vowel Onset Points (VOPs) is important for recognition of utterances of Consonant-Vowel (CV) subword units using the neural network models and for spotting the subword units in continuous speech. For CV utterances, VOP is the point at which the consonant ends and the vowel begins. This point is characterized by a rapid change in the speech signal parameters such as energy, linear prediction residual energy and spectral flatness. In this paper, we show that a neural network can be trained to capture the characteristics of the speech signal at the VOPs. The trained neural network can then be used to detect the VOPs of the CV utterances. It is demonstrated that this method gives a significantly better (by 11%) performance compared to the energy derivative based method.

1 Introduction

One approach to develop a vocabulary independent continuous speech recognition system is to spot the subword units in continuous speech. In Indian languages, the Consonant-Vowel (CV) units occur with high frequency. The CVs are the basic units of speech production and capture the significant coarticulation effects in these units. For spotting the CV units in continuous speech, it is necessary to identify the segments that contain CV units and then use a classifier to decide the CV classes of the segments. We consider a method based on detection of Vowel Onset Points (VOPs) to identify the CV unit segments in continuous speech. The number of CV classes is large (145) because each of the 29 consonants can be followed by any of the 5 vowels to form a CV unit. Because of the large number of classes and high similarity among several CV classes, it is necessary to use pattern classification models capable of good discrimination.

Multi Layer Perceptron (MLP) based models have been commonly used for complex pattern classification tasks. The MLP models are suitable for classification of fixed length patterns. However, durations of the utterances vary not only for different CV classes, but also for a particular CV class. Therefore the patterns representing the speech signal parameters of the varying duration utterances of CV units will be of varying length. It is necessary to develop a method for representing the CV utterances by fixed length patterns. The important information necessary for classification of CV utterances is present in the region around the VOP. Therefore a fixed duration (200 msec) segment around the VOP can be processed to extract a fixed length pattern to represent a CV utterance [1]. Portions of a CV utterance in the beginning and the end that are not included in the fixed duration segment contain only redundant information that can be ignored. In this approach, it is necessary to develop a method for detection of VOPs in CV utterances with a good accuracy. Therefore, detection of VOPs is important for developing MLP based classifiers for CV utterances and for spotting the CV units in continuous speech [2]. The method for detection of VOPs has to be robust and speaker-independent. In this paper, we address the issues in detection of VOPs in CV utterances.

In section 2, we present two methods for detection of VOPs. The first method is based on identifying the point at which the positive derivative of the short-time energy is maximum. The second method that we propose uses a neural network trained to detect the VOPs. In section 3, we present the studies on detection of VOPs using these two methods and compare their performance.

2 Methods for Detection of VOPs

In this section, we briefly review the methods proposed earlier for detection of VOPs and then present the methods used in our studies. The method proposed in [3] identifies the points at which there is a rapid increase in the vowel strength as VOPs. The vowel strength is computed using the difference in 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 analyzed to obtain the amplitude spectrum and compute the vowel strength. This method requires unvoiced/voiced classification of the speech signal. Two methods proposed in [4] and [5] 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 [6]. 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 and ratio of the front and back 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. In the proposed method, it is not necessary to segment the speech signal into unvoiced/voiced regions or vowel-like/nonvowel-like regions.

The first method that we consider for detection of the VOP in a CV utterance is to identify the point at which there is a rapid increase in the energy of the speech signal. This is the point at which normally the consonant ends and the vowel begins in a CV unit. This point is detected by computing the derivative of the energy and locating the point at which the positive derivative is maximum. This method is found to be suitable for CVs where the consonants belong to the category of unaspirated stop consonants. However, for the CVs where the consonants belong to the other categories (aspirated stop consonants, fricatives, nasals, semivowels, and affricates), the energy derivative based method is not robust. It is mainly because the rapid increase in the energy also occurs at the point at which the consonant of a CV unit starts after the initial silence. Therefore the energy derivative at VOP may not be the maximum. This leads to erroneous detection of VOPs.

Parametric analysis of the speech signal for CV utterances has shown that three parameters, namely, (1) signal energy, (2) linear prediction (LP) residual energy and (3) spectral flatness, are significantly different in the regions immediately before and after the VOPs. The signal energy and the LP residual energy parameters show a rapid increase at the VOPs whereas the spectral flatness parameter shows a decrease.

We train a multilayer perceptron network to detect the VOPs by using the trends in the speech signal parameters at the VOPs. The network used for detecting the VOPs consists of 2 hidden layers with 25 nodes in the first hidden layer and 20 nodes in the second hidden layer. The input layer of the network contains 9 nodes and the output layer has 3 nodes. Though the main function of the network is to detect the VOPs, it should also be trained to minimize the number of false alarms in the detection of VOPs. Therefore 3 nodes are used in the output layer of the network with one node (labeled as VOP node) to indicate the presence of VOPs and the other two nodes (labeled as Pre-VOP and Post-VOP) to indicate the absence of VOPs.

For training the network to detect the VOPs, it is necessary to give the values of the above three parameters extracted from a speech signal frame immediately before the VOP and a frame immediately after the VOP. We also give the ratios of the parameters in these two frames because the trend in these parameters contains the information for detection of the VOPs. Thus the signal energy, residual energy and spectral flatness parameters extracted from two frames around the VOP and the ratios of the parameters in the two frames are used to form a vector of 9 values that is input to the network during training. One of the frames starts at 15 ms before the VOP and the other frame starts at 5ms after the VOP. The duration of frames is 10 ms. For training the network, the vectors are extracted from CV utterances for which the VOPs are manually identified. The desired outputs are specified to give the maximum value (1.0) for the output node labeled as VOP and the minimum value (0.0) for the other two output nodes labeled as Pre-VOP and Post-VOP. We also extract two other vectors from each CV utterance. One vector is derived from two frames in the region before the VOP. One of these frames starts at 35 ms and the other at 15 ms before the VOP. The desired outputs for this training vector specify the maximum value for the Pre-VOP node and minimum values for the VOP and Post-VOP nodes. Another vector is derived from two frames in the region after the VOP. One of these frames starts at 35 ms and the other at 55 ms after the VOP. The desired outputs for this training vector specify the maximum value for the Post-VOP node and minimum values for the VOP and Pre-VOP nodes.

For detection of the VOP in a CV utterance using the network trained as above, a 9-dimensional parameter vector extracted at every 10 ms 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 at 20ms 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 hypothesized as the VOP of the CV utterance. This method can also be extended for detection of the VOPs of multiple CV segments in the continuous speech signal of a sentence. In the next section, we compare the performance of the energy derivative based method with the neural network based method.

3 Studies on Detection of VOPs

Isolated utterance data of 145 CV classes was collected from two male speakers. For each class, 12 tokens were collected from each speaker. The actual VOP has been manually marked for every utterance so that the performance of the methods for detection of VOPs can be studied. The training data of the neural network model consists of the 9-dimensional parametric vectors extracted from 4 tokens of each CV class from one speaker (called as Speaker 1). The test data set includes remaining 8 tokens of each class from speaker 1 and 12 tokens of each class from speaker 2. For each CV utterance, the VOP hypothesized by a method is determined as explained in the previous section. The difference between the hypothesized VOP, denoted as H, and the actual VOP, denoted as A, is computed to give the deviation $(A - H)$ in 10 millisecond units . The performance of the energy derivative based method and the neural network based method in detection of the VOPs of CVs is given in Table 1. The performance is given as a percentage of the test data for which the VOPs are detected with a particular value of deviation. The performance is given for the deviation upto 30 msec that may be acceptable for classification of CV patterns. It is noted that the neural network based method gives a significantly improved performance by about 11%.

The performance of the methods for different categories of CVs is analyzed. The percentage of test utterances of a category of CVs for which the VOPs are detected with a particular deviation value is shown in a distribution plot in Fig.1. The dashed line corresponds to the energy derivative based method, and the solid line corresponds to the neural network based method. Each plot shows the performance for different values of deviation, $(A - H)$. The percentage shown for devia-

Table 1: Performance of the energy derivative based method and the neural network based method in detection of VOPs. The performance is given as a percentage of the test data for which the VOPs are detected with a particular value of deviation.

Deviation ($A-H$) in msec	Energy Derivative Method	Neural Network Method
0	26.5	27.8
+10	12.8	6.2
-10	17.1	33.8
+20	2.4	1.3
-20	11.0	12.1
+30	2.0	1.6
-30	6.7	5.2
Total	78.3	89.1

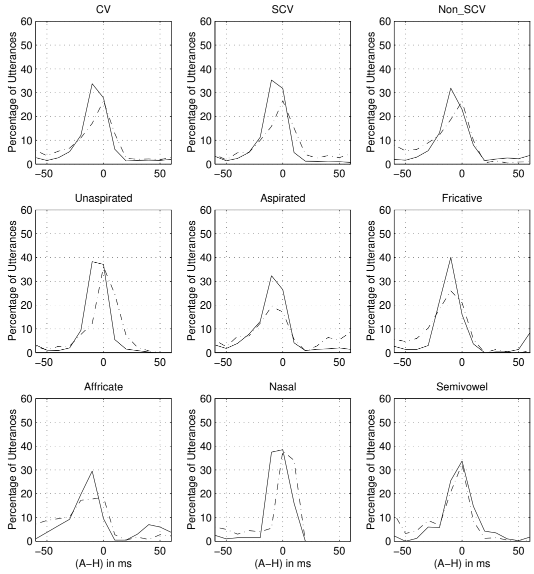

Figure 1: Distribution plots showing the performance of the energy derivative based method (dashed line) and the neural network based method (solid line) in detection of VOPs for different categories of CVs. The performance is given as a percentage of the test utterances of a category of CVs for which the VOPs are detected with a particular value of deviation.

tion of 60 msec corresponds to the utterances for which the VOP is detected with a deviation equal to or greater than 60 msec. The plots are shown for the following categories of CVs: (a) All CVs, (b) Stop_Consonant-Vowels (SCVs), (c) Non_SCVs, (d) Unaspirated SCVs, (e) Aspirated SCVs, (f) Fricative CVs, (g) Africate CVs, (h) Nasal CVs and (i) Semivowel CVs. The plots show that the neural network based method gives an improved performance for each category of CVs. The performance of the methods is found to be poor for aspirated SCVs and semivowel CVs.

A performance measure, D, to quantify the performance of the methods in detection of VOPs is defined as follows:

$$D = \sum_{i=-6}^{+6} W_i F_i \tag{1}$$

where $i = (A-H)/10$, $W_i = |i|$ is the weight factor associated with i, and F_i is the percentage of the utterances for which the VOPs are detected with the deviation corresponding to i. The value of the performance measure D for the two methods and for different categories of CVs is given in Table 2. The high value for D indicates that the performance is poor. It is observed that the value of D is significantly smaller for the neural network based method compared to that for the energy derivative based method. It is also noted that even the neural network method performs poorly for the aspirated SCVs.

The method used for detection of the VOP in a CV utterance by scanning the speech signal using a trained neural network can be extended for detection of VOPs of multiple CV segments present in the utterance of a word or a sentence. The waveform of the speech signal and the output for the VOP node (with 11-point hamming window smoothing) of the neural network are plotted in Fig.2 for the utterance /na:ga:rjunasa:gar/. It is seen that the locally dominant peaks in the output plot correspond to the VOPs of CV segments in the utterance. There may be spurious peaks that do not correspond to the VOPs. It is necessary

Table 2: The value of the performance measure D for the energy derivative based method and the neural network based method, for different category of CVs.

Category of CVs	Energy Derivative Method	Neural Network Method
All CVs	191.03	147.48
SCVs	186.31	131.75
Non_SCVs	196.85	166.85
Unaspirated	116.50	103.12
Aspirated	256.12	160.38
Fricative	190.00	183.00
Affricate	247.50	238.50
Nasal	134.00	87.00
Semivowel	182.75	123.00

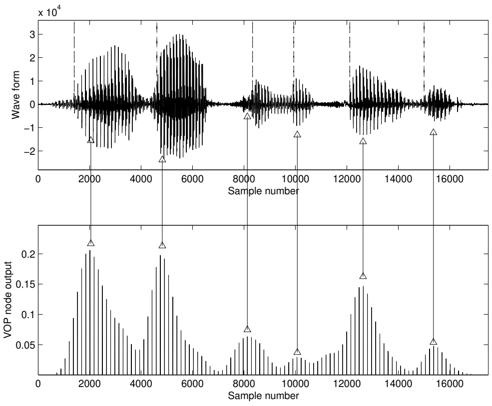

Figure 2: Plots of the waveform of the speech signal and the output of the neural network model trained to detect the VOPs, for the utterance /na:ga:rjunasa:gar/. The actual VOPs (dashed lines) and hypothesized VOPs (lines with arrows) are shown.

to eliminate the spurious peaks and also ensure that the VOPs of all the CV segments are detected so that the method can be used for spotting the CV segments in continuous speech.

4 Conclusions

In this paper, we have addressed the issues in detection of vowel onset points. We have considered the energy derivative based method and a neural network based method. It has been demonstrated that the neural network based method gives a significantly better performance. It is necessary to study the effect of deviation of the hypothesized VOP from the actual VOP on the classification performance of the CV recognition system. It is also necessary to develop the methods for detection of the VOPs in utterances with multiple CV segments so that the corresponding subword units can be spotted in continuous speech.

References

[1] C.Chandra Sekhar, *Neural Network Models for Recognition of Stop Consonant-Vowel (SCV) Segments in Continuous Speech.* PhD thesis, Indian Institute of Technology, Madras, 1996.

[2] C. Chandra Sekhar and B. Yegnanarayana, "Neural network models for spotting Stop Consonant-Vowel (SCV) segments in continuous speech," in *Proceedings of International Conference on Neural Networks*, pp. 2003–2008, 1996.

[3] D. J. Hermes, "Vowel-onset detection," *Journal of the Acoustical Society of America*, vol. 87, pp. 866–873, February 1990.

[4] A. Bendiksen and K. Steiglitz, "Neural networks for voiced/unvoiced speech classification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 521–524, April 1990.

[5] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 250–255, April 1993.

[6] H. Kasuya and H. Wakita, "An approach to segmenting speech into vowel- and nonvowel-like intervals," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 27, pp. 319–327, 1979.