Detection of Vowel Onset Points in Continuous Speech using Autoassociative 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

Detection of vowel onset points (VOPs) is important for spotting subword units in continuous speech. For consonant-vowel (CV) utterances, VOP is the instant at which the consonant part ends and the vowel part begins. Accurate detection of VOPs is important for recognition of CV units in continuous speech. In this paper, we propose an approach for detection of VOPs using autoassociative neural network (AANN) models. A pair of AANN models are trained for each CV class to capture the characteristics of speech signal in the consonant and vowel regions of that class. The trained AANN models are then used to detect VOPs in continuous speech. The results of studies show that the proposed approach leads to significantly less number of spurious hypotheses.

1. Introduction

Speech recognition involves transformation of 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. One approach to develop a vocabulary independent continuous speech recognition system is to spot the subword units in continuous speech. Most production and perception units of speech sounds are dynamic in nature. 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. Therefore, we consider CV units of speech as the basic units for speech recognition. In Indian languages, the CV units occur with high frequency. 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.

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 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. In this approach, it is necessary to develop a method for detection of VOP in CV utterances with a good accuracy. Therefore, detection of VOPs is important for developing MLP or SVM based classifiers for CV utterances and for spotting the CV units in continuous speech.

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 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]. 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.

The paper is organized as follows: In Section 2, we briefly review the VOP detection methods. Then we propose a method for detection of VOPs using AANN models in Section 3. We illustrate the behavior of the proposed method in Section 4. In Section 5, we study the performance of VOP detection methods.

2. Methods for detection of VOPs

In this section, we review some of the existing methods for detection of VOPs in continuous speech. 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 method proposed in [3] classifies the speech signal into voiced/unvoiced/silence regions using a neural network classifier, and then labels the voiced regions as vowel and nonvowel regions. Segmentation of continuous speech into vowel-like and nonvowel-like regions was proposed in [4]. 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 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. 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. The MLFFNN classifier is trained using the vectors extracted from the three different regions of each utterance. For detection of VOPs in continuous speech using the trained MLFFNN model, a 9-dimension 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 at 20 msec after this point. Thus the continuous speech signal of an utterance is scanned by the network to detect the VOPs. The output of the network indicates the strength of evidence for presence of the VOP at that point in a continuous speech. The locally dominant peaks in the outputs of the VOP node correspond to the VOPs of CV segments. The VOP node output is smoothened using a 11-point Hamming window. A peak picking method is used to hypothesise the VOPs at locally dominant peaks. In the next section, we propose a method for detection of VOPs in continuous speech utterances using AANN models.

3. AANN model based approach for detection of VOPs

Let us consider the five layer AANN model shown in Fig. 1, which has three hidden layers. The processing units in the first and third hidden layers are nonlinear, and the units in the second hidden layer can be linear or nonlinear. As the error between the actual and the desired output vectors is minimised, the cluster of points in the input space determine the shape of the hypersurface obtained by the projection onto the lower dimension space [1]. The distribution of clusters will be captured by the AANN models.

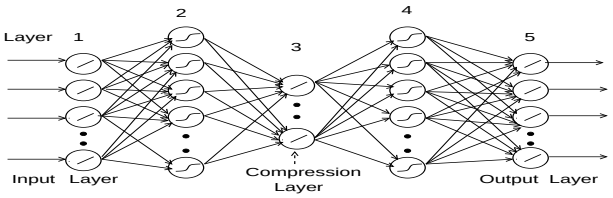

Figure 1: Five layer AANN model.

For each CV class, one AANN model is developed for capturing the distribution of feature vectors in the consonant region and another model for the vowel region of the utterances of that CV class. The distribution is expected to be different for the consonant and vowel regions of a class. The implementation of VOP detection system using AANN models is explained in the following subsections.

3.1. Speech database and representation

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, energy, their first order derivatives and their second order derivatives. Thus the dimension of each frame is 39. The VOP detection systems are developed for the data of each of the three languages.

3.2. Development of AANN models

For each CV class, two AANN models 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

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

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 number of units in layers 2 and 4 are chosen empirically. The use of four units in the middle layer is based on the study reported in [1]. The activation function for the nonlinear units is a hyperbolic tangent function. The network is trained using error backpropagation algorithm in pattern mode.

3.3. VOP detection using AANNs

Short-time analysis of continuous speech utterance is performed using a frame of size 20 msec with a shift of 5 msec resulting in a sequence of frames. One frame at a time is given to the pairs of AANN models of all the CV classes. The evidence E of a model for the current frame $\mathbf{x}$ is defined as:

$$E = \exp\left(-\frac{\| \mathbf{x} - \mathbf{y} \|^2}{\| \mathbf{x} \|^2}\right) \quad (1)$$

where $\mathbf{x}$ is the input vector given to the model and $\mathbf{y}$ is the output vector given by the model. Let $E_c(k)$ and $E_v(k)$ be the evidence obtained from consonant and vowel region models of k^{th} class respectively. Then the hypothesised region $H(k)$ of the current frame by the models of class k is given by:

$$H(k) = \begin{cases} C & : \quad if \quad E_c(k) \geq E_v(k) \\ V & : \quad otherwise \end{cases}$$

Let N be the total number of CV classes. The models of each of these classes hypothesise the current frame as a consonant or a vowel frame. Let N_c and N_v be the number of hypotheses as belonging to the consonant region and the vowel region, respectively. Then, the current frame is hypothesised as belonging to the consonant region or vowel region as follows:

$$H = \begin{cases} C & : \quad if \quad N_c \geq N_v \\ V & : \quad otherwise \end{cases}$$

This method is used to obtain a sequence of labels for a sequence of frames of continuous speech utterance. From this sequence of labels the VOP frames are hypothesised as follows: If the current frame is labeled as C and there are at least eight successive frames labeled as V, then the fourth frame to the right of the current frame is identified as the VOP frame. The block diagram of the proposed system for detection of VOPs in continuous speech utterances is shown in Fig. 2.

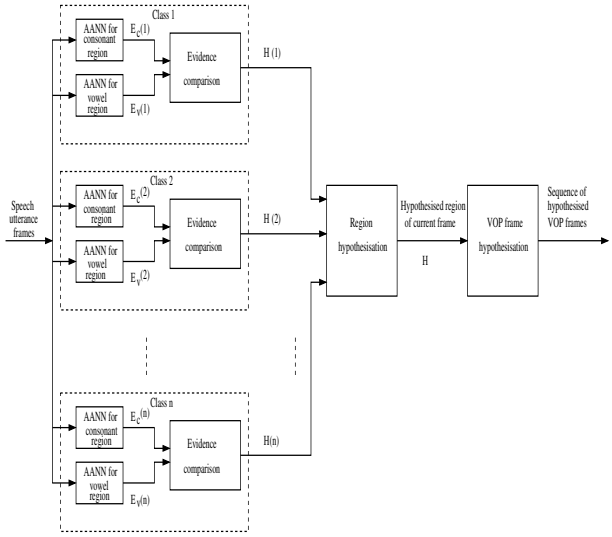

Figure 2: Block diagram of the proposed system for detection of VOPs in continuous speech.

4. Detection of VOPs in continuous speech

For illustration, we consider a Tamil language speech utterance /kArgil pahudiyilirundu UDuruvalkAraraL/ 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. 3 (a). The hypothesised region labels using the proposed system are shown in Fig. 3 (b). Using the procedure described in Section 3, 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. 3 (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. 3 (d).

It is seen that there are three VOPs (their sample numbers are indicated in boldface) that have been missed around the locations 4920, 12080, and 12520 corresponding to the syllables /hu/, /Du/, and /ru/, respectively. It is seen that there are fewer than 8 vowel region hypotheses around these locations. The VOP at location 15480 is not with in 25 msec of its actual location. Therefore, it is hypothesised as spurious VOP. In the next section, the performance of the VOP detection methods is compared for a large number of sentences in three Indian languages.

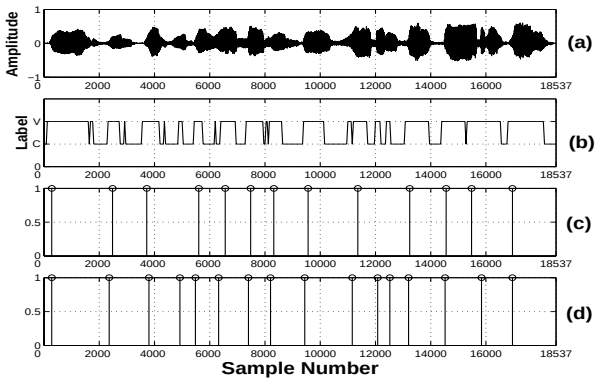

Figure 3: Plots of the (a) Waveform of the speech signal, (b) Hypothesised region labels for each frame, (c) Hypothesised VOPs, and (d) Manually marked (actual) VOPs for the Tamil language sentence /kArgil pahudiyilirundu UDuruvalkArarhaL/.

5. Studies on detection of VOPs

In this section we study the performance of VOP detection method based on AANN models. For comparison we consider the method based on MLFFNN model described in Section 2. The performance is measured in terms of number of matching, missing and spurious VOPs of speech utterances. 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 CV 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 methods explained in Sections 2 and 3. The performance of the VOP detection methods for each of the languages is given in Table 2. 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 the actual VOP, the VOPs of such segments are considered as the missing hypotheses. When there are multiple hypotheses with in 25 msec around the actual VOP or the hypothesised VOP does not fall in this range, such hypotheses are considered as spurious ones. The performance of VOP detection methods for the data of three languages is given in Table 2. The average performance for different methods is given in 3.

It is seen from Table 3 that the performance of both the methods is nearly the same for matching case. However, the VOP detection method based on AANN provides significantly less number of spurious VOPs. The missing VOPs in case of AANN based method are mainly for CV units with consonants as semivowels, fricatives and nasals.

Table 2: Performance for detection of VOPs in continuous speech for each of three languages.

Language	Method	Matching	Missing	Spurious
Tamil	MLFFNN	1217	363	353
	AANN	1059	321	92
Telugu	MLFFNN	1000	648	650
	AANN	1131	517	96
Hindi	MLFFNN	483	213	296
	AANN	486	210	73
Total	MLFFNN	2700	1224	1299
	AANN	2676	1248	261

Table 3: Comparison of the average performance 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	Matching	Missing	Spurious
MLFFNN	68.80	31.19	33.10
AANN	68.19	31.80	6.65

6. Summary and conclusions

In this paper we have proposed a method for detection of vowel onset points (VOPs) for consonant-vowel (CV) units in continuous speech using autoassociative neural network (AANN) models. This method can be used for detection of VOPs in utterances with multiple subword units so that the corresponding subword unit can be spotted in continuous speech. It has been demonstrated that the AANN based method gives significantly less number of spurious hypotheses. It is necessary to study the effect of deviation of hypothesised VOP from the actual VOP on the classification performance of the CV recognition system. It is also necessary to reduce the number of misses.

7. 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] 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] 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.

[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.