Proceedings of
ICISIP - 2005

Vowel Onset Point Based Variable Frame Rate Analysis for Speech Recognition

A. Nayeemulla Khan, B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, India.
email: {nayeem, yegna}@cs.iitm.ernet.in

Abstract

The consonant and transition regions around the vowel onset point bear important contextual cues for distinguishing between different classes of consonant vowel units. It is important to model these regions for improved speech recognition rates. Conventional fixed frame rate analysis does not distinguish between the different regions of the speech signal. In this paper we detail an approach to emphasize the consonant and transition regions in speech sounds using a variable frame rate, anchored around the vowel onset point. The usefulness of this technique is demonstrated on an isolated consonant-vowel recognition task for Indian languages.

1. INTRODUCTION

Consonant-vowel (CV) units are among some of the most difficult sound units to recognise. Due to the similarities in their speech production mechanism the confusability among the CV units is high. A CV unit may have some or all the following significant events manifesting in the acoustic signal; closure, burst, aspiration, transition and vowel. Some of these events manifest for short durations. Observing the spectrogram of these units, it is noticed that the characteristics of the burst region is different from that of the transition region. The transition region carries information about the preceding consonant, which is useful in discriminating between different CV units. A better representation of consonant and the transition regions of speech would improve the performance of the recognisers.

In conventional speech processing systems, the speech signal is windowed into frames of 20 to 30msec duration, and features extracted using a frame shift of 10 to 15 msec. In such fixed frame rate (FFR) processing, no importance is attached to the different regions of speech which carry important cues for perception and classification of speech sounds. Variable frame rate analysis is used to emphasize areas where the spectral characteristics change rapidly [1][2][3].

The approach commonly used in variable frame rate (VFR) analysis is to first extract the feature vectors based on a fixed frame rate. Then measure the change in spectral characteristics between adjacent frames. Based on some limiting criteria choose to keep the frames or drop them. If the change is high then more features may be extracted in this region. In [1] the Euclidean distance between the current frame and the last retained frame is determined, and this is compared to the threshold. The current frame is dropped if the distance is below the threshold. In [4] the norm of the first derivative of the feature vector is used to come up with the decision. An entropy based approach in [5] can also be used to assist in the decision making. A review of the different approaches to VFR analysis is detailed in [2]. There they show that VFR techniques are useful in the presence of additive noise at low signal to-noise ratio. They further conclude that it does not improve the results over FFR analysis for clean speech.

The information present in the transition region needs to be suitably represented for correct classification of CV units [6]. It is important to suitably model the dynamics of the consonant and the CV transition region. In this paper we propose an approach wherein VFR analysis anchored around the vowel onset point (VOP) is used to improve the recognition rate of isolated consonant vowel units.

The reminder of the paper is organised as follows. In the next section we discuss about the VOP. Section 3 deals with VFR analysis. In Section 4 we describe the dataset used for the study. The recognition studies conducted are explained in 5. We summarise the study in Section 6.

2. VOWEL ONSET POINT

Figure 1 shows a typical CV unit. The end of the consonant part and the beginning of the vowel part signifies that VOP. Since the vowel region is prominent in the signal due to the large amplitude characteristics and due to the periodicity of excitation, it is easier to locate the event compared to other speech production events. The VOP can be used as an anchor point for different tasks like, extracting fixed length patterns useful for neural network classifiers and spotting CV segments in continuous speech. Acoustic-phonetic description and the acoustic cues useful for detection of the VOP for different CV units is described in [7]. An algorithm for detection of VOP can be found in [8]. The closure, burst and aspiration region of the CV unit are present before the VOP in any stop consonant vowel. The consonant region lies before the VOP. The transition and the vowel regions are present after the VOP. A fixed duration of the signal around the VOP contains most of the information necessary for classification of CV units [9].

0-7803-8840-2/05/$20.00 ©2005 IEEE

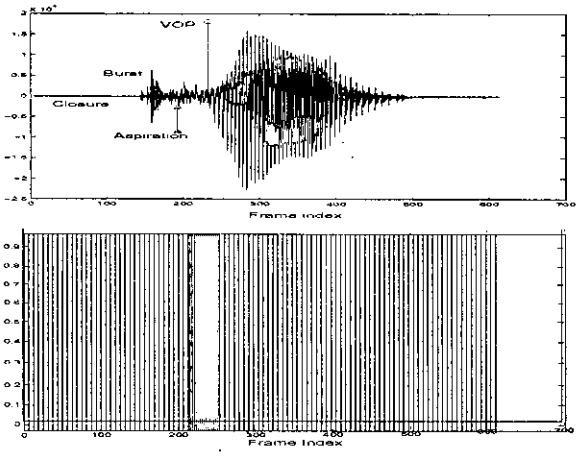

Fig. 1: A typical CV unit /kha/ (top), Typical frame selection for VFR analysis (bot.).

3. VFR ANALYSIS

Earlier studies suggest that in steady regions frames can be dropped and in regions of rapid changes more frames at closer intervals can be used to represent the signal [4]. Once the VOP is identified accurately the transition region immediately after the VOP can be determined. The consonant region lies before the VOP. It is one of the main information bearing elements of speech. The duration of most consonants is around 15 to 30 msec. For aspirated sounds it may be larger as in Figure 1. The transition region is around 20 to 30 msec after the VOP. In the region of 20 msec before and 20 msec after the VOP the dynamics of the speech signal varies rapidly. Features are extracted for every 15 msec frame shifted by 5 msec for the entire CV utterance except for 40 msec duration around the VOP. For this region the frame shift is reduced to 1 msec. This procedures attempts to capture the dynamic variation in the region around the VOP while using the standard processing for the rest of the signal.

In the studies conducted frames in the vowel region are not dropped, as they do not seem to improve the performance of the recogniser for the isolated CV recognition task. The use of VFR is dependent on the detection of the VOP, as the region of interest is derived based on the correct location of the VOP.

4. DATABASE

Indian languages are basically syllabic in nature. The basic sounds in all these languages are of the CV type, a consonant followed by a vowel. There are about 29 consonants and 5 vowels totaling 145 basic sound units. The details of CV units in Hindi can be found in [10]. Speech data for isolated utterances of each of the 145 basic CV units is collected at 16 kHz sampling frequency. The data was collected for 12 repetitions of each unit from each of three speakers. The total number of utterances are $3 \times 145 \times 12 = 5,220$. Ten utterances of each speaker are used for training and two for testing. The test set contains $3 \times 145 \times 2 = 870$ utterances. The VOP's are located manually for each of the CV utterances.

TABLE 1: PERFORMANCE OF THE CV RECOGNITION SYSTEM.

Analysis type	Recognition rate (%)
FFR system	83.6
VFR system	90.6

5. EXPERIMENTAL DETAILS

From each CV utterance for every 15 msec frame with a frame shift of 5 msec, 13 cepstral coefficients along with their first and second derivatives forming a 39 dimensional feature vector is extracted. A HMM based isolated CV recogniser that recognises the test utterance as one among the CV units in the vocabulary is constructed. Each CV unit is modelled by a standard left-to-right no skip HMM with 5 emitting states and a single Gaussian per state. The models are trained in an isolated word fashion. This system formed the baseline system. The performance of this system for the test set is 83% as shown in Table 1. This is the highest performance reported for this database. Earlier studies on this database using constrain satisfaction neural networks [11] and multilayer feed forward neural network [10] report 60% and 61% recognition rates respectively.

For a region of 20 msec before and 20 msec after the VOP the features are extracted at every 1 msec interval emphasizing the consonant and transition region where the dynamics of the signal changes rapidly. Features resulting from the VFR analysis is used in training HMM models of the same structure as previously. The performance of the VFR based CV recognition system is reported in Table 1. The improvement in performance is a significant 8.4% over the baseline fixed frame rate system.

A. Error in location of the VOP

In the previous section the VOP's used were derived manually. It is difficult to obtain manually marked VOP's for large datasets. Automatic approaches to VOP detection are detailed in [7][12][13]. In all these approaches there exists a margin of error. In the VFR analysis as the region where the features are extracted at a finer rate is based on the VOP, the performance of the system is thus dependent on the accurate location of the VOP. The effect of errors in location of the VOP were studied. The errors in location of the VOP is simulated assuming the VOP is located either 10 or 20 msec to the left or right of the manually marked VOP. When the error in location of the VOP is 20 msec to the left of the VOP (-20 msec) the region of emphasis (40 msec around the VOP) is entirely in the consonant region, and the transition region is dealt with by FFR analysis. If the error in location of the VOP is 20 msec to the right of the VOP (+20 msec), the converse is true. That is the region of emphasis is entirely in the transition and vowel region of the CV unit and the consonant region is handled by FFR analysis. The performance of the CV unit recogniser for the different errors is shown in Table 2.

TABLE 2: PERFORMANCE OF THE CV RECOGNITION SYSTEM IN THE PRESENCE OF ERRORS IN VOP LOCATION.

Error in location of VOP (msec)	Recognition rate (%)
-10	91.1
-20	90.5
+10	89.4
+20	84.7

From Table 2 we see that if the error in location of the VOP is to the left of the actual VOP, more frames are extracted from the consonant region and the performance does not vary much. Whereas if the error in the VOP location is to the right of the actual VOP then most of the consonant region is left out, and the extra frames are extracted only from the transition and vowel regions leading to a performance closer but marginally better that the baseline system. This result reinforces the fact that the consonant region along with the transition region carries important information useful for CV classification.

6. SUMMARY AND DISCUSSION

In the study we have shown that VFR analysis anchored around the VOP is useful in improving the performance of the CV recognition system. Further the system is tolerant to errors of around 20 msec in location of the VOP from the actual position. This study needs to be extended to continuous speech. In continuous speech the error in VOP detection by automatic means is generally higher and the procedure also gives rise to spurious VOP's. If the spurious VOP's are large in number, it will lead to extracting a large number of frames around them which are extraneous. One drawback in a HMM based system is that, if the system encounters too many frames then the insertion error would increase dramatically. This issue needs to be addressed while extending the study to continuous speech.

REFERENCES

[1] S. M. Peeling and K. M. Peeling, "Variable frame rate analysis in the ARM continuous speech recognition system," *Speech Comm.*, vol. 10, pp. 155–162, June 1991.

[2] J. Macías-Guarasa, J. O. ñez, J. M. Montero, J. Ferreiros, R. Córdoba, and L. F. D. Haro, "Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition," in *Proc. EUROSPEECH* , (Geneva, Switzerland), pp. 1809–1812, Sept. 2003.

[3] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 3, (Istanbul, Turkey), pp. 1783–1786, June 2000.

[4] P. L. Cerf and D. V. Compernolle, "A new variable frame rate analysis method for speech recognition," *IEEE Signal Processing Letters*, vol. 1, pp. 185–187, Dec. 1994.

[5] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable fram rate analysis of speech signals and its application to ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, (Quebec, Canada), pp. 549–552, May 2004.

[6] K. N. Stevens, "Models for production and acoustics of stop consonants," *Speech Comm.*, vol. 13, pp. 367–375, Dec. 1993.

[7] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Proc. Signal Proc. Comm.*, (Indian Inst. Sci., Bangalore, India), pp. 81–88, July 2001.

[8] S. R. M. Prasanna, *Event based analysis of speech.* PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Jan. 2004.

[9] C. C. 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, Apr. 1996.

[10] S. V. Gangashetty, K. S. Rao, A. N. Khan, C. C. Sekhar, and B. Yegnanarayana, "Combining evidence from multiple modular networks for recognition of consonant-vowel units of speech," in *Proc. IEEE Int. Joint Conf. Neural Networks*, vol. 1, (Portland, Oregon, USA), pp. 686–691, July 2003.

[11] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Constraint satisfaction model for enhancement of evidence in recognition of consonant-vowel utterances," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 3, (Maryland, USA), pp. 201–204, July 2003.

[12] J. Y. S. R. K. Rao, "Recognition of Consonant-vowel (CV) utterances using modular neural network models," MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, May. 2000.

[13] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Detection of vowel onset points in continuous speech using autoassociative neural network models," in *Int. Conf. Spoken Language Processing (INTERSPEECH 2004 - ICSLP)*, (Jeju Island, Korea), 2004.