Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Identify Vowels and Consonants of Punjabi Speech Using SVM

Parwinder Kaur*, Amanpreet Kaur
Department of Computer Science & Engineering
India

Abstract-- Speech consists of acoustic pressure waves created by the voluntary movements of anatomical structures in the human speech production system. These waveforms are broadly classified into voiced and unvoiced speech. Voiced sounds (vowels for example), produce quasi-periodic pulses of air which are acoustically filtered as they propagate through the vocal tract. The main distinction between vowels and consonants is that vowels resonate in the throat. Formants are exactly the resonant frequencies of a vocal tract when pronouncing a vowel. This paper addresses the issues in segmentation of Punjabi speech into sub-word units of speech using Formants and support vector machines (SVMs). Many studies have been conducted to identify and discriminate vowels and consonants using acoustic/articulator differences. In this study the Punjabi speech is segmented into smaller speech units and each unit is classified either consonant or vowel using the Formant frequencies. This process when further combined with recognition of each unit will form a complete speech recognition system. The proposed detection strategy is tested with the speech signals recorded from the television broadcast.

Keyword-- Support Vector Machine (SVM), Formants Frequencies, Speech recognition system, Punjabi speech.

I. INTRODUCTION

In phonetics the basic units of speech are vowels and consonants. All the languages contain the both kinds of phonemes and always it is hard to draw a line dividing these two categories. Vowels can be characterized as periodic sounds produced with the vibration of vocal cords and the airflow from the lungs is not blocked. Consonants are often non-periodic sounds produced with the obstruction of airflow from the lungs and with or without vocal cords vibration. Vowels form the nuclei of the syllables whereas the consonants form the onset and coda. Formants are exactly the resonant frequencies of a vocal tract when pronouncing a vowel. A Formant is a concentration of acoustic energy around a particular frequency in a speech wave**.** In other words, these are the meaningful & distinguishable frequency components. There are several formants, each at a different frequency, roughly one in each 1000 Hz band. Each formant corresponds to a resonance in the vocal tract. The information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. The formant with the lowest frequency is called f_1, the second f_2, and the third f_3. Most often the two first formants, f_1 and f_2, are enough to disambiguate the vowel. These two formants determine the quality of vowels in terms of the open/close and front/back dimensions (which have traditionally, though not entirely accurately, been associated with the position of the tongue). Thus the first formant f_1 has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as [i] or [u]); and the second formant f_2 has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as[u]). Vowels will almost always have four or more distinguishable formants; sometimes there are more than six. However, the first two formants are most important in determining vowel quality, and this is often displayed in terms of a plot of the first formant against the second formant, though this is not sufficient to capture some aspects of vowel quality. The main objective of this paper is to detect and segment Punjabi speech signal into a sequence of consonant and vowel units. The proposed algorithm is composed of three stages, as shown in Fig.1. In the first stage, the input audio is segmented into 20ms-long frames with a 5ms shift, where formant frequencies for each frame is calculated. In order to group the frames into V/C in phoneme level, a silence detection algorithm using spectral centroid and signal energy is proposed. In the second stage the formant frequencies for each frame is calculated using the Linear prediction analysis. In the third stage each frame is identified as either vowel or consonant using the support vector machine.

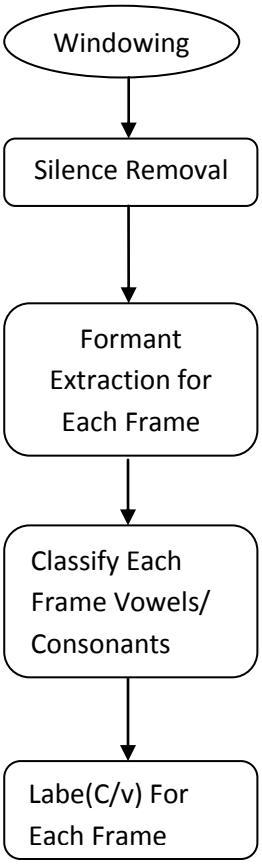

Fig 1. Diagram of the stages in the proposed algorithm

II. SUPPORT VECTOR MACHINE

Support vectors are the data points that lie closest to the decision surface
• They are the most difficult to classify
• They have direct bearing on the optimum location of the decision surface
• We can show that the optimal hyperplane stems from the function class with the lowest "capacity" (VC dimension).
Support vector machines (SVMs) have been shown to give a good generalization performance in solving pattern recognition problems. The main idea of a support vector machine for pattern classification is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples of a class is maximized. In this study a Non-linear support vector machine with Gaussian Kernel is used for classification. In a Non-linear classifier the dot product is replaced by a non- linear kernel function. Three nonlinear kernel functions considered in our studies are as follows:
Sigmoidal kernel -
$K(X, Xi) = \tanh(0.001XT\ Xi - 1.0)$
Polynomial kernel -
$K(X, Xi) = (XT\ Xi + 1)2$
Gaussian kernel -
$K(X, Xi) = \exp(-\left| X - Xi \right| 2 / 0.01)$

In this study Formant frequencies are used as Feature vector to train and test the support vector machine for classifying Consonant or Vowel. Formants are the distinguishing or meaningful frequency components of human speech and of singing. The vowels are represented purely quantitatively by the frequency content of the vowel sounds so that human can better distinguish between various vowels. The formant with the lowest frequency is called f1, the second f2, and the third f3. Most often the two first formants, f1 and f2, are enough to disambiguate the vowel. These two formants determine the quality of vowels and thus the first formant f1 has a higher frequency for an open vowel and a lower frequency for a close vowel and the second formant f2 has a higher frequency for a front vowel and a lower frequency for a back vowel. Vowels will almost always have four or more distinguishable formants; sometimes there are more than six.

III. FORMANT ESTIMATION USING LPC ANALYSIS

There are two methods for estimating formants from the predictor parameters. The widely used and the method used in this study for formant analysis is factoring the predictor polynomial and based on the roots obtained formants are extracted. The other method is to obtain the spectrum and choose the formants by a peak picking method. Initially the predictor order f is chosen using the formula given below.

f= round (fs/1000) +2,
where fs is sampling frequency in Hz

After performing the LPC analysis on the speech signal, to identify the missing fundamental frequency auto correlation is performed on the speech signal. Autocorrelation is the cross-correlation of a signal with itself. Autocorrelation is a widely used tool for finding repeating patterns, such as the presence of a periodic signal which has been buried under noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies.

R (i) = E{x (n) x (n-i)},
where R is the auto correlation

IV. SILENCE REMOVAL

Silence removal can be considered as a one of the efficient dimensionality reduction technique in speech signal processing. Therefore in this study as a pre-processing step a silence removal method is applied. The two simple audio features namely the signal energy and the spectral centroid are used for silence removal from the speech signal. Initially the two feature sequences are extracted from the given input speech signal and thresholds are determined for each sequence. Speech segments are detected based on the simple thresholding technique.

V. CLASSIFICATION

After extracting formant frequencies for each frame of the input signal each frame is classified into vowel or consonant using the Non-linear support vector machine. Initially the support vector machine is trained with a hand labeled speech corpus in which the vowel and consonant segments are marked manually. The classified smaller speech units are finally labeled automatically as either Consonant or Vowel. So the final output is a sequence of labels whose length is equal to the number of frames of the speech signal.

VI. CONCLUSION

This study presented a simple technique for segmentation of Punjabi speech signal in to Consonant and vowel units using formants. The proposed method is easier to implement and the results are also promising that this approach can be applied in other speech processing application. Further this research can be extended to build a speech recognition system based on the individual Consonant/Vowel unit.

REFERENCES

[1] K. Amanpreet, and S. Tarandeep, "Segmentation of Continuous Punjabi Speech Signal into Syllables," Proceedings of the World Congress on Engineering and Computer Science 2010 Vol. I, WCECS 2010, San Francisco, USA, October 20-22, 2010.

[2] S. Nishi, and S. Parminder, "Automatic Segmentation of Wave File", International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 267-270.

[3] G Lakshmi Sarada, et al. "Automatic transcription of continuous speech into syllable-like units for Indian languages", Sadhana, Vol. 34, Part 2, April 2009, pp. 221–233.

[4] C. Vimala and V.Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal, Vol. 2, No.1, 2012, pp. 1-7.

[5] Prica Biljana et al. "Recognition of vowels in Continuous Speech by using Formants", Facta Universitatis, Vol. 23, No. 3, December 2010, pp. 379-393.

[6] Anwar Jamil Muhammad et al. "Automatic Arabic Speech Segmentation System", International Journal of Information Technology", Vol. 12, No. 6, 2006, pp. 102-111.

[7] Natarajan Anantha V. et al. "Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies", International Journal of Computer Applications, Vol. 56, No.15, October 2012, pp. 24-27.

[8] Rao Preeti et al. "Speech formant frequency estimation: evaluating a nonstationary analysis method", Elsevier, 2000, pp. 1655-1667.

[9] Rahman Mijanaur Md. et al. "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches", International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, 2012,pp. 131-138.

[10] Greenberg S. "Strategies for Automatic multi-tier annotation of spoken language corpora", Proceedings of the 8th European Conference on Speech Communication and Technology, 2003, pp. 45–48.

[11] Sharma M. & Mammone R. "Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge", ICSLP Proceedings, Vol. 2, 1996, pp. 1237-1240.

[12] Tolba F. M. et al. "A Novel Method for Arabic Consonant/Vowel Segmentation using Wavelet Transform", International Journal of Intelligent Computing and Information Sciences, Vol. 5, No. 2, 2005, pp. 353-364.

