Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies

V. Anantha Natarajan
Dept. of Computer Science & Engineering
Annamalai University
Annamalai Nagar-608002, India.

S. Jothilakshmi
Dept. of Computer Science & Engineering
Annamalai University
Annamalai Nagar-608002, India.

ABSTRACT

This paper addresses the issues in segmentation of continuous 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/articulatory differences. In this study the continuous 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

1. 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 [1]. 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.

The speech recognition system transforms the given input speech signal in to corresponding Text which carries the same message as the speech signal. The major two steps involved in continuous speech recognition is identifying the phonetic units and recognizing each unit to produce the text output. Another important problem analyzed in the area of speech recognition is vowel onset point (VOP) detection. Vowel onset detection is the location where the onset of vowel utterance takes place. VOPs' are used to identify the Syllable boundaries in the continuous speech signal. An approach for detection of VOPs using auto associative neural network (AANN) models is proposed by Suryakanth V. Gangashetty & et al., in [2]. Formants are exactly the resonant frequencies of a vocal tract when pronouncing a vowel. Using the formant frequencies an attempt to carry out Vowel Recognition in Serbian language is presented in [3]. The main objective of this paper is to detect and segment continuous speech signal into a sequence of consonant and vowel units. After detecting the basic speech units each unit can be recognized using any trained classifier to form a speech recognition system.

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 signanl 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.

The rest of the paper is organized as follows: the section 2 presents the various background concepts related to this study. The section 3 describes the silence removal process using the signal energy and the spectral centroid. The section 4 presents an overview of the formant extraction process using Linear predictive coding technique. The section 5 and section 6 describes the classification and the experiments & discussions respectively. Finally the section 7 concludes the results of this study.

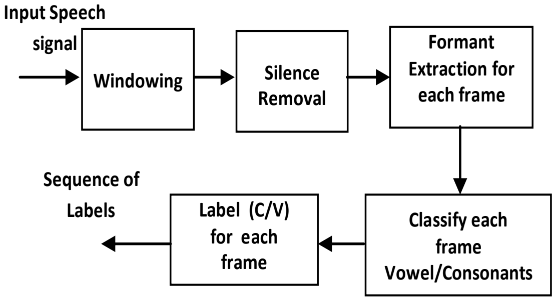
Input Speech signal
Windowing
Silence Removal
Formant Extraction for each frame
Classify each frame Vowel/Consonants
Label (C/V) for each frame
Sequence of Labels

Fig 1. Block Diagram of the stages in the proposed algorithm

2. BACKGROUND CONCEPTS

Support vector machines (SVMs) have been shown to give a good generalization performance in solving pattern recognition problems [7]. 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 [8]. 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. The Fig.2 shows the process of mapping a non-linearly separable function into a high dimension linear separable function. Three nonlinear kernel functions considered in our studies are as follows:

Sigmoidal kernel -

$\boldsymbol{K(X , X_i) = \tanh (0.001X^T X_i – 1.0)}$

Polynomial kernel -

$\boldsymbol{K(X , X_i) = (X^T X_i + 1)^2}$

Gaussian kernel -

$\boldsymbol{K(X , X_i) = \exp(-| X - X_i |^2 / 0.01)}$

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 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 and thus the first formant f_1 has a higher frequency for an open vowel and a lower frequency for a close vowel and the second formant f_2 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. In this study, five formant frequencies for each frame are calculated.

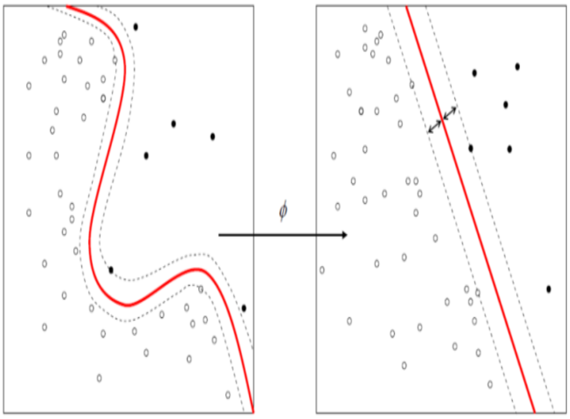

Fig 2. A non-linearly seperable functions into a higher dimension linearly seperable function.

3. SILENCE REMOVAL ALGORITHM

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 [4]. 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.

Signal Energy , Energy of the ith frame ,

$$E(i) = \frac{1}{N} \sum_{n=1}^{N} |x_i(n)|^2$$

where N is the length of the sample .

Spectral centroid is given by

$$C_i = \frac{\sum_{k=1}^{N} (k+1)\, x_i\,(k)}{\sum_{k=1}^{N} x_i\,(k)} \,.\, x_i\,(k)$$

where X$_i$(k) is the DFT of the ith frame.

The energy and the spectral centroid will be low in the silent region of the speech signal. In Fig.2 the waveform (a) shows the short time energy plot of the original and filtered signal. The waveform (b) shows the plot of the Spectral centroid and the waveform (c) shows the waveform of the original signal.

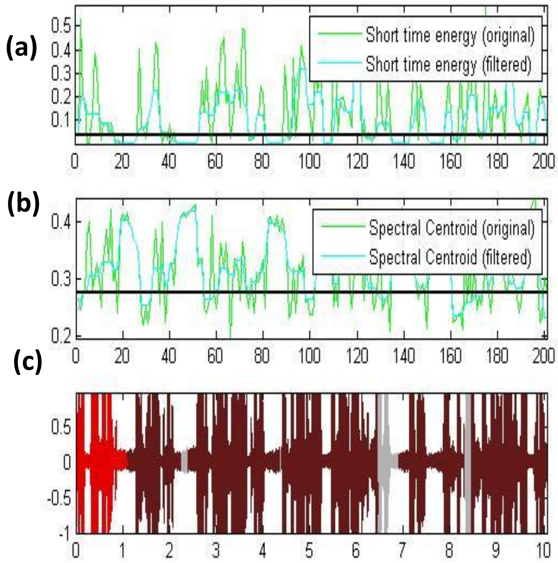

Fig 3. Waveforms showing the plot of Signal energy and Spectral centroid

The Fig. 3 shows the obtained results after pre-processing stage in the proposed algorithm. The waveform labeled (a) presents the Short time energy of the original and filtered speech signals. The waveform (b) shows the spectral centroid of the original and filtered speech signals. The waveform (c) represents the original speech signal with silent regions being highlighted.

4. 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 [5]. The other method is to obtain the spectrum and choose the formants by a peak picking method. Initially the predictor order p is chosen using the formula given below.

p = round(fs/1000)+2 ,

where fs is sampling frequency in Hz

The Linear predictor is given by,

$$\hat{x}(n) = \sum_{i=1}^{p} a_i \, x(n-i)$$

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*

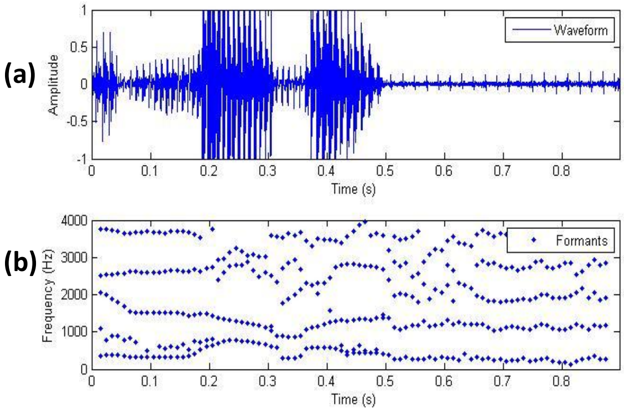

Fig 4. (a) Speech wave form a Tamil word utterance (b) Formants Frequency Curve

The Fig.4 describes the result of the Formant estimation process using Linear predictive coding. The Figure shows the Frequency curve for a short Tamil word utterance by a male speaker.

5. 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.

6. EXPERIMENTS & RESULTS

The Tamil language speech samples used for training and testing the Support vector machine are recorded from the television broadcasts containing both male and female speaker. Each speech corpus is sampled at the rate of 8kHz and the total duration of the speech samples is 1hour. Each training sample is processed using the silence removal algorithm and then windowed using a hamming window for the short-time analysis of the speech signal. Each frame is analyzed using Linear predictive coding for estimating formant frequencies. Along with the formant frequencies the consonant or vowel label is given as input to the support vector machine during training process. During testing time the test signal is processed as described above and given to the trained support vector machine. For each frame of the test sample the SVM produces either 0 or 1 whether the frame belongs to Consonant or vowel respectively. Then finally each frame is labeled depending upon the output of the SVM.

Kernel used	Accuracy in %
Sigmoidal	59.8
Polynomial	65.4
Gaussian	83.46

Table 1. Classification accuracy of Speech units

The Table 1 shows the classification accuracy of the various kernels used for classification. Of all the kernels used in this study the Gaussian Kernel shows better performance in terms of classification accuracy. The label sequence generated as the final output describes the class of the each frame in the speech signal. To identify the boundary of each syllable a rule based approach can be used or the onset of the vowel will give an idea about the boundary of the individual units

7. CONCLUSION

This study presented a simple technique for segmentation of continuous 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. The experiments and results show that the Gaussian kernel gives better accuracy when compared to other kernels used in this study. Further this research can be extended to build a speech recognition system based on the individual Consonant/Vowel unit.

8. REFERENCES

[1] B. van Ooyen, A. Cutler and D. Norris. Detection of Vowels and Consonants with Minimal Acoustic Variation. *Speech Communication,* vol. 11, 1992.

[2] Suryakanth V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana. Detection of Vowel Onset Points in Continuous Speech using Autoassociative Neural Network Models. In: Proc of Int. Conf. Spoken Language Processing , INTERSPEECH 2004-ICSLP . Page(s):1081-1084.

[3] Biljana Prica and Siniˇsa Ili´c. Recognition of Vowels in Continuous Speech by Using Formants. SER.: ELEC. ENERG. vol. 23, no. 3, December 2010, Page(s): 379-393

[4] Theodorous Giannakopoulous. A method for silence removal and segmentation of speech signals. Computational Intelligence Laboratory (CIL), Institute of Informatics and telecommunications, NSCR Demokritos, Greece 2009.

[5] G. Palshikar. Simple algorithms for peak detection in time-series, trddc technical report'09. In Technical Report, TRDDC.

[6] Ali A., Bhatti.S, Mian M.S. Formants Based Analysis for Speech Recognition, Engineering of Intelligent Systems, 2006 IEEE International Conference , Page(s): 1 – 3.

[7] C.J.C.Burges, \A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, Page(s): 121-167,1998.

[8] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.