SPEECH SIGNAL ENDPOINT DETECTION USING HIDDEN MARKOV MODELS

M. Masroor Ahmed [1], Abdul Manan Bin Ahmed [1], Razib M. Othman & Sheraz Khan[2]
[1]Faculty of Computer Science & Information System
Universiti Technologi Malaysia
[2]Département Traitement du Signal et Systèmes Électroniques,
Supelec, Paris, France
Email: masroorahmed@gmail.com , manan@fsksm.utm.my, razib@fsksm.utm.my
sheraz.khan@supelec.fr

Abstract:

A major cause of errors in automatic speech recognition system is the inaccurate detection of the beginning and ending boundaries of test and reference patterns. Separation of speech and silence segments in automatic speech recognition algorithms occupies a fundamental position. The improper demarcation of these segments reduces system's efficiency, since; the system has to execute processing on the portion of segments which are not needed. A comprehensive evaluation of ASR systems showed that more than half of the recognition errors are caused due to wrong word boundary detection [1][2] Therefore the desired characteristics for an endpoint detection are reliability, robustness, accuracy, adaptation, simplicity and real time processing etc[3]. This paper discusses a robust algorithm for the detection of speech and silence segments in an input speech signal based on Hidden Markov Models

Keywords: Hidden Markov Models, Automatic Speech Recognition, SNR, Endpoint

1.0 Introduction

Distinguishing speech and non speech segments in a digital speech signal is the ultimate objective in an automatic speech recognition system. Due to its fundamental importance it has assumed the role of a key preprocessing step in speech signal processing. It cannot be denied that accurate extraction of speech and silence segments always play a dominant role in enhancing the accuracy and efficiency of speech recognition system. The obvious reason for this improved performance is that the system saves time in exercising un-necessary execution on non speech segments. Generally, the system with wrong endpoint detection suffers with two negative effects, i.e. Recognition errors because of the incorrect boundaries; and secondly increased computations "wasted" on the non-speech events in the utterance. The importance of accurate extraction of speech and silence segments can not be ignored and this goal can be achieved by thoroughly examining the speech signal [1].

This is general observations that for recording a speech signal or for using an ASR system we can not find an ideal environment which is completely noise free. The environments include many

high level or non stationary noises, such as in the mobile phone with speech command system. The understandable sources for these noises may include speakers lip smacks, mouth clicks, environment door slams, fans, machines and transmission channel noise, cross talk. The variability of durations and amplitudes for different sounds makes reliable speech detection difficult [5]. Extraction of speech and non speech segments is quite simple in noise free environment or environment where there is minimal amount of noise is present. Signal short-time energy, zero-crossing rate or spectral energy is frequently employed in traditional endpoint detection algorithms. These features proved effective in clean environment but they fail in maintaining the efficiency and accuracy of the speech signals with noise [2][3]

Since a real time speech signal is time signal, therefore, to extract the important information from the input speech signal, this is of fundamental importance that we know various frequency components present in the input signal. For this purpose, generally Fourier transforms are applied. The apparent shortcoming of this transform is that, this is unable to provide any information of the spectrum changes with respect to time. This concept assumes that the signal is stationary. To overcome this deficiency, a modified method-short time Fourier transform allows representing the signal in both time and frequency domain through time windowing functions. The window length determines a constant time and frequency resolution. We will be able to get good information about the input signal if, a shorter time windowing is used in order to capture the transient behavior of a signal [6].

In this paper, the goal is to develop an HMM based algorithm for accurate detection of speech and silence segments. An algorithm which also guarantees the important characteristics like reliability, robustness, accuracy, adaptation, simplicity, and real-time processing [4][5]. For the implementation of algorithm in question, Discrete Hidden Markov Models were opted due to various reasons [7]:

- Static modeling technology is a cheap alternative both in training cost and computational complexity, but in return it produces worse recognition results.
- Neural Network approach showed the highest resources demand. Therefore, it is less suitable for developing applications
- Continuous Hidden Markov Models requires the extensive computation for every Markov state in the system. Therefore, like Neural Network, it is also resource hungry.

2 Input Data

For training the system dataset consisting of first name, last name and complete name was used. It contains samples from different speakers which covers the speech variability and pronunciation problems. The speech corpus is sampled at 22050 Hz with short silence before and after each utterance. The signal is transformed into Mel scale coefficients. The Mel scale coefficients as extracted features are selected because they imitate into some extent the feature selection into human ear.

Since the left – right topology provides a good traversal from beginning state to the

ending state, therefore, it is followed for the implementation of the algorithm.

3 Training an HMM for Speech and Silence

For this purpose, two models have been trained. One is responsible for detecting the speech segment whereas; the other serves the purpose of marking silent areas in the input signal. Speech and silence models are initialized using random numbers. For training and testing Mel Frequency Cepstrum Coefficients (MFCC) was extracted and a frame size of 500 points was taken while training HMM, whereas, a frame size of 100 points was taken while testing the system. In addition to it, all the silence portions from the training signal are removed. This step significantly contributes in preparing and training a reliable speech model.

Since the speech and silence models were initialized with random numbers, therefore, the re-estimation of all the parameters of HMM is needed. This re-estimation is done with the help of Baum-Welch re-estimation formulae.

The Baum-Welch algorithm for (re)estimating HMM parameters $\lambda = (A, B, \pi)$ with a given set of speech training data {O}. The values of λ will be chosen so that the likelihood P (O | λ) is maximized for the given {O}.

$$\xi(i,j) = \Pr(i_t = q_i, i_{t+1} = q_j \mid O, \lambda)$$

i.e the probability of a path being in state q_i at time t and making a transition to state q_j at time t+1, given the observation sequence and the model. The re-estimation formula for π, A, B are:

- $\bar{\pi} = \gamma_t(i) \qquad 1 \le i \le N$
- $\bar{a}_{ij} = \dfrac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$
- $\bar{b}_j(k) = \dfrac{\sum_{\substack{t=1 \\ O_t = k}}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)}$

Here $\bar{\pi}$, $\bar{a}_{ij}$ and $\bar{b}_j(k)$ are the estimated values of π, a_{ij} $b_j(k)$ of the HMM model The γ terms will be calculated efficiently with a "forward-backward algorithm". In essence, this algorithm makes an efficient organization of the many different state sequences into step-wise transitions.

For identifying speech or silence segments from the input signal, the algorithm concentrates on the values P(O | λ) independently, both for speech and silence. A segment is considered ‘silence’ when its P(O | λ) is greater than the probability of observation given lambda of speech segment; otherwise, the same segment is labeled as a speech segment.

3.1 Decoding an HMM

For evaluating an HMM forward-bacward algorithm is used, but it does not calculates the best state sequence, which is the best path. Exhaustive search is expensive since it absorbs all the computer resources. Therefore, it is needed that a mechanism should be there which can efficiently calculate the best path. Viterbi algorithm suits well in this situation[8].

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the Viterbi path – that result in a sequence of observed events, especially in the context of hidden markov models. The forward algorithm is a closely related algorithm for computing the probability of a sequence of observed events. The algorithm makes a number of assumptions. First, both the observed events and hidden events must be in a sequence. This sequence often corresponds to time. Second, these two sequences need to be aligned, and an observed event needs to correspond to exactly one hidden event. Third, computing the most likely hidden sequence up to a certain point t must depend only on the observed event at point t, and the most likely sequence at point $t - 1$. These assumptions are all satisfied in a first-order hidden Markov model [8][9].

3.2 Maximization over the Model

Given the initial state sequence, we maximize over the model

$$\max_{\lambda} \; P(x, \lambda \mid q)$$

The maximization entails estimating the model parameters from the observation given the state sequence. Estimation is performed using the Baum-Welch re-estimation procedure[8].

4 Results

The algorithm has been tested rigorously and it has produced very appreciative results. The results are shown in the following figures in which speech and silence segments are identified and are marked properly.

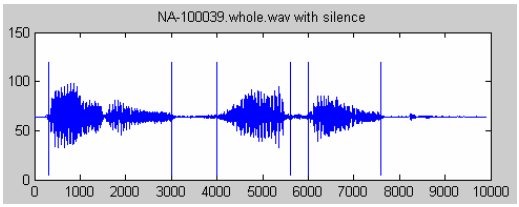

Figure 1

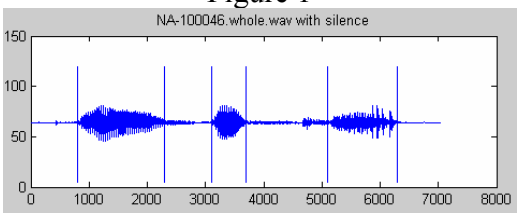

Figure 2

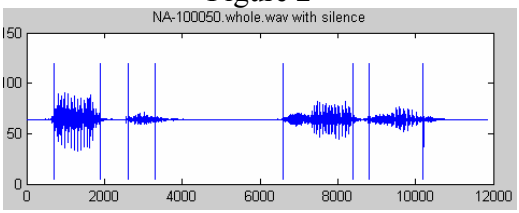

Figure 3

4.1 Calculating Accuracy

The percentage of accuracy is calculated with the help of following relation [10].

(Manual Labeling – Auto Labeling / Manual Labeling) * 100

The goal of the speech segment extraction is to separate acoustic events of interests from other parts of the signal i.e. the silence and the background. This is an important front end processor. Figure 4 shows a comparison between various endpoint detection algorithm and the proposed method.

The recognition rate of many spoken commands ASR systems is heavily dependent on proper extraction of endpoint of a speech signal.

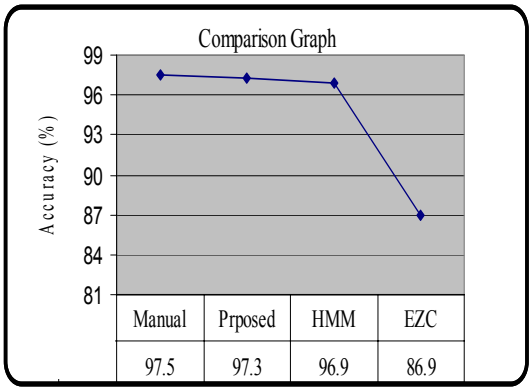

Figure4: Comparison with the Existing Methods

When these points are neatly identified, then the recognition rate would be higher. Thus a goal for discovering a new and sophisticated technique that can play a decisive role in protecting system degradation is successfully achieved. The unique feature of the proposed method is its design on HMM foundations.

From Figure 4, it is clear that energy and zero crossing (EZC) method has lowest accuracy, i.e. 86.9, the accuracy of HMM based endpoint detection method is higher as compared to EZC, i.e. 96.9 [10] and the accuracy of the proposed method is higher than EZC and HMM based methods, i.e. its 97.3.

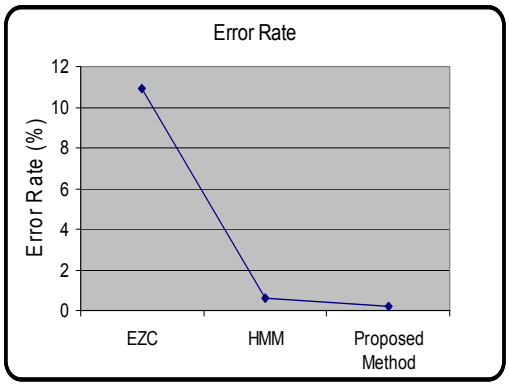

Figure 5: Comparison of Error Rate

The accuracy is calculated by manually labeling the speech signals and making it a yardstick for checking other available methods. The recognition rate achieved due to hand labeling is 97.5. So this technique is taken as reference method for drawing comparison with any of the subsequent technique.

The overall recognition error rate shown in Figure 5 caused due to energy and zero crossing method is 10.9 %, i.e. (97.5 - 86.9 / 97.5), the HMM based method contributes 0.6 % of that error, the hand labeling method causes. Whereas, the proposed technique contributes only 0.2 % of the recognition error rate.

5 Conclusions

This paper aimed to highlight the importance for the accurate extraction of speech and silence segments from a speech signal. The distinctive feature described in this paper is the efficient use of HMMs for maintaining and improving the performance of an automatic speech recognition system. The technique studied in this research produced appreciative results. One limitation of the technique is that, it is capable of handling the noise free signal, but this can be extended to deal with the noisy signals at various SNR levels in continuous enviornment .

References:

[1] Wilpon J.G. and Rabiner L.R. (1987)Application of hidden Markov models to automatic speech endpoint detection. *Comput. Speech Language*, 2:321-341

[2] Lamel, L., Rabiner, L., Rosenberg, A.and Wilpon, J.;(1981). An improved endpoint detector for isolated word

recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing , Volume: 29 , Issue: 4 , Aug 1981*

[3] Junqua J.C., Mak B. and Reaves B.,(1994) A Robust Algorithm for Word Boundary Detection in the Presence of Noise. *IEEE Trans. on Speech and Audio Processing.*

[4] Liang-Sheng Huang and Chung-Ho Yang (2000). A novel approach to robust speech endpoint detection in car environments *ICASSP '00. Proceedings. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000., Volume: 3 , 5-9 June 2000*

[5] Qi Li, Jinsong Zheng, Tsai, A. and Qiru Zhou. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition, *IEEE Transactions on Speech and Audio Processing , Volume: 10 , Issue: 3 , March 2002* : 146 – 157

[6] Bou-Ghazale, S.E.; Assaleh, K.;(2002). A robust endpoint detection of speech for noisy environments with application to automatic speech recognition *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02).*

[7] Bourlard, H., Morgan, N. and Renals, S. (1992) Neural nets and hidden Markov models: Review and generalizations, *Speech*

[8] Rabiner, L.R.;(1989). A tutorial on hidden Markov models and selected applications in speech recognition Proceedings of the *IEEE , Volume: 77 , Issue: 2 , Feb. 1989*

[9] http://wikipedia.org

[10] Abdullah,W.H. (2002). HMM-Based techniques for speech segments extraction. *Journal of Scientific Programming, vol.10, no. 3, pp.221-239.*