SPEECH RECOGNITION IN NOISY ENVIRONMENT USING DIFFERENT FEATURE EXTRACTION TECHNIQUES

M. G. Sumithra¹, M. S. Ramya² and K. Thanuskodi³

Bannari Amman Institute of Technology, Sathyamangalam, India
₁mgsumithra@rediffmail.com, remiraju15@gmail.com
²Akshaya College of Engg. and Tech., Coimbatore, India
E-mail: ³thanush_dr@rediffmail.com

Abstract: In this paper, different feature extraction methods for speech recognition system such as Mel-frequency cepstral coefficients (MFCC), linear predictive coefficient cepstrum (LPCC) and Bark frequency cepstral coefficients (BFCC) are implemented and the comparison is done based on average recognition accuracy. We suggest a noise robust isolated word speech recognition system which can be applied in various noisy environments. In this method, Kalman filter is used to remove the background noise and to enhance the speech signal. The enhanced signal is integrated into the front end of recognition system in order to guarantee high performance in noisy environment conditions. Performance of different recognition methods are compared based on recognition accuracy rate in noisy environment.

Keywords: Speech Enhancement, Feature Extraction Methods, Dynamic Time Warping, VQ\HMM

I. INTRODUCTION

Automatic speech recognition (ASR) is the process of determining a sequence of words spoken by human using machines [1]. The goal of ASR is to have speech as a medium of interaction between man and machine and it is desired that an ASR system is robust to these unwanted variabilities. Speech can get affected by the channel through which it passes or the environment in which it is produced and/or recorded. Speech recognizers used in clean conditions normally perform poorly in noisy environments, due to the mismatch between training and testing data [2]. As recognition spoken language technologies are being transferred to real word applications, the need for greater robustness against noisy environments is becoming increasingly apparent. However, real word conditions differ from ideal or laboratory conditions, causing mismatch between training and testing and consequently, inducing performance degradation in the automatic speech recognition systems. A simple special case of mismatch situation is encountered when the testing signal is corrupted by various additive noises while the training data are clean. In order to solve the noise degradation problem of the speech recognition, many front-end processing techniques, speech enhancement have to be used [3]. After enhancing, the noisy speech signal is used for testing phase in recognition system.

In this paper, kalman filter is used to enhance the noisy utterances. Different feature extraction methods are used for analysis in DTW based speech recognition system. From the analysis best extraction method will be used for recognition system. In recognition part, DTW is used to detect the nearest recorded voice [4]. In testing phase clean speech signal is added with different type of noises in different SNR values. The average recognition accuracy of different extraction methods are calculated and compared.

This paper is organized as follows: In Section2, the proposed method is explained. Section3 deals with speech enhancement and explains about Kalman filter. Section4 includes different feature extraction methods. Section5 gives detail about template \ model creation. A discussion of the results obtained with this method is given in Section6. Section7 outlines the conclusion and future work.

II. PROPOSED METHOD

To develop a noise robust recognition system [5], testing signal is added with different type of noises and enhanced using kalman filter. The features of enhanced signal is extracted and used for matching
with trained model. The proposed method comprises of analysis of different feature extraction methods and performance comparison of different recognition methods. MFCC, LPCC and BFCC are the feature extraction methods are implemented and their average recognition accuracy rate is compared. Based on results, best extraction method will be preferred for noise robust recognition system. DTW and VQ\HMM are different recognition methods are used and their average recognition accuracy rate is calculated in noisy environment.

III. SPEECH ENHANCEMENT

In automatic speech recognition systems, the performance degrades badly in the case of adverse environments with very low SNR. It was found that the recognition rate can be improved by applying a speech enhancement algorithm to the degraded speech to improve the quality and intelligibility. The problem of cleaning noisy speech still poses a challenge to the area of signal processing. For enhancing the noisy speech, Kalman filter [7] was chosen.

A. Kalman Filter

From the literature, it is understood that noise reduction is seen more in the enhanced signal by using Kalman filter. Hence Kalman filter was selected as a preprocessor for enhancing the speech in this work.

Kalman filtering has been shown to be an effective speech enhancement technique, in which a speech signal is usually modeled by an autoregressive (AR) process and represented in the state-space domain [8]. The advantage of Kalman filtering, as compared to spectral suppression, is that it can completely overcome the tonal noise problem and achieve quite good speech quality by reducing the processing distortion introduced to a speech signal. Kalman filtering not only uses the statistical characteristics of signal and noise, but also utilizes the well-known autoregressive (AR) speech model which has been proven to be effective for modeling the human speech production system.

The Kalman filter has two distinct phases: Predict and Update. The predict phase uses the state estimate from the previous time step to produce an estimate of the state at the current time step. This predicted state estimate is also known as the a priori state estimate because, although it is an estimate of the state at the current time step, it does not include observation information from the current time step. In the update phase, the current a priori prediction is combined with current observation information to refine the state estimate. This improved estimate is termed the a posteriori state estimate.

IV. FEATURE EXTRACTION METHODS

A. Mel-Frequency Cepstral Coefficient

Mel-Frequency Cepstral Coefficient (MFCC) vectors are used to provide an estimate of the vocal tract filter [9]. Background noise energy level is evaluated at the beginning and the end of speech signal and energy thresholds are applied to find speech beginning and end points. The preemphasized speech signal is blocked into frames of $N = 256$ samples, with adjacent frames separated by $M = 128$ samples. Then windowing is done to minimize the speech signal discontinuities at the beginning and end of each analysis frame.

After windowing, FFT is applied. Then a spectrum is passed through 20 mel scale triangular filter bank. The mel scale is a critical band frequency scale that takes into account the frequency perception in the human auditory system. DCT is applied to the log Mel scale filter outputs and thus 12 mel-frequency Cepstral coefficients (MFCC) are obtained. Then Cepstral filtering is performed.

$$\text{Mel}(f) = 2595 \log \left( 1 + \frac{f}{700} \right)$$  \hspace{1cm} (1)

B. Linear Prediction Cepstral Coefficient

Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech [9]. Linear predictive analysis provides both an accurate estimate of the speech parameters and also an efficient computational model of speech. The LP filter contains the vocal tract information by describing the spectral
envelope of the speech. The LP filter is an all-pole filter of order \( p \) as given by

\[
H(z) = \frac{1}{1 - \sum_{k=1}^{p} d(k)z^{-k}}
\]  

(2)

The filter or predictor coefficients \( d(k) \) are denoted by the \( p \)-dimensional vector \( d \). In order to compute the LP filter coefficient vector \( d \), the auto-correlation of the signal \( t(n) \) is first computed as :

\[
R_i = \sum_{n=0}^{N-1-i} t(n)t(n+i)
\]  

(3)

where \( N \) is the number of samples in the signal and \( i \) is the time lag of the autocorrelation function.

C. Bark Frequency Cepstral Coefficient

Instead of using Mel filter bank, Bark filter bank has been applied and equal loudness pre-emphasis with intensity to loudness power law has been applied to the MFCC like features [9]. BFCC is processing the spectra and cosine transforms to get the cepstral coefficients. Only first thirteen cepstral features of each windowed frame of speech utterances were taken.

\[
Bark = 13\arctan(0.00076 f) + 3.5\arctan((\frac{f}{7500}))
\]  

(4)

V. CREATION OF TEMPLATE \ MODEL

A. Dynamic Time Warping

The Dynamic Time Warping (DTW) distance measure is a technique that has long been known in speech recognition community. It allows a non-linear mapping of one signal to another by minimizing the distance between the two [4].

The dynamic time warping algorithm creates an alignment between two sequences of feature vectors, \( (T_1, T_2 ... T_N) \) and \( (S_1, S_2, ..., S_M) \). A distance \( D(i, j) \) can be evaluated between any two feature vectors \( Ti \) and \( Sj \). This distance is referred to as the local distance. In DTW the global distance \( D(i, j) \) of any two feature vectors \( Ti \) and \( Sj \) is computed recursively by adding its local distance \( d(i, j) \) to the evaluated global distance for the best predecessor.

The best predecessor is the one that gives the minimum global distance \( D(i, j) \) at row \( i \) and column \( j \):

\[
D(i, j) = \min_{m,k} [D(m,k)] + d(i, j)
\]  

(5)

The computational complexity can be reduced by imposing constraints that prevent the selection of sequences that cannot be optimal. Global constraints affect the maximal overall stretching or compression. Local constraints affect the set of predecessors from which the best predecessor is chosen. Dynamic Time Warping (DTW) is used to establish a time scale alignment between two patterns. It results in a time warping vector \( w \), describing the time alignment of segments of the two signals. It assigns a certain segment of the source signal to each of a set of regularly spaced synthesis instants in the target signal. During DTW pattern matching test pattern is matched to each reference template and global dissimilarity measure is calculated. A reference template providing minimal global distance measure is accepted as recognized word by the decision logic. DTW approach uses the Nearest Neighbor (NN) decision rule. The distance scores for all the reference patterns are sent to a decision rule which gives the word with least distance as recognized word. The distance measure between two feature vectors is calculated using Euclidean distance metric

\[
D_q = \frac{1}{p} \sum_{k=1}^{p} (T_{MFCC}(i,k) - R_{MFCC}(j,k))^2
\]  

(6)

B. Vector Quantization(VQ)

Vector Quantization is the method of automatically partitioning feature space in to different clusters based on training data. For speech recognition using vector quantization one codebook is used for each word of the recognition vocabulary. The VQ codebook is a discrete representation of speech. We will generate the codebook by using LBG algorithm. This algorithm has used two times, one is on training time and other is testing time. In training time we will generate codebook for delta cepstrum coefficients. In testing we will use stored codebook for getting the indices of codebook that give minimum distortion.

The VQ will find a codebook index corresponding to the vector that best represents a given spectral vector for an input vector sequence \( V[v(1), v(2), v(3), ..., v(N)] \). VQ will calculate the vector distance between each vector in codebook \( C[c(1), c(2), c(3), ..., c(P)] \) and each vector \( v(n) \) and the codebook index with minimum distance
will be chosen as output. After VQ, a sequence of codebook indexes $I \{I(1), I(2), I(3), ..., I(N)\}$ will be produced. The vector distance between an input vector $v(n)$ and each vector in codebook are calculated as follows

$$Distance(p) = \sum_{k=1}^{K} [v(n)(k) - c(p)(k)]$$

$$i(n) = \text{agr } \min_p (\text{Dis } tan ce(p))$$

C. Hidden Markov Model (HMM)

HMM is typically an interconnected group of states that are assumed to emit a new feature vector for each frame according to an emission probability density function associated with that state. HMM is used to represent the utterance of the word and to calculate the probability of that the model which created the sequence of vectors. Given the observation sequence $O = O_1, O_2, ..., O_T$ and the model $\lambda = (A, B, \pi)$, how we compute $Pr(O|\lambda)$, the probability of the observation sequence. The evaluation problem is efficient way of computing this probability using forward and backward algorithm. The decoding helps to find the best matching state sequence given an observation sequence and solve using Viterbi algorithm. The model parameters $\lambda = (A, B, \pi)$ to maximize $Pr(O|\lambda)$ resolved by using Baum-Welch algorithm [11].

for initialization,

$$\delta_1 (j) = \log \pi_j + \log b_j(o_1)$$

for termination,

$$p(o|\lambda) = \max [\delta_T (i)] \text{ for } t = T$$

The acronym used in the algorithm:

$N$ is Number of states

$T$ is Number of frames for feature vectors $o = [o_1, o_2, ..., o_T]$

$A_{ij}$ is State transition between $i$ and $j$

$A$ is their $N$-by-$N$,

$B = \{ \log b_j(o_t) \}$ is a $N$-by-$T$ matrix in log output probability, and

$\delta_t (j)$ is the likelihood value at the time index $t$ and state $j$.

VI. RESULTS AND DISCUSSION

The proposed algorithm is implemented using MATLAB 10. Database is collected by pronouncing the words from 0-9 by 20 speakers. Each word was pronounced ten times by each speaker in noise-free conditions, yielding 200 utterances per speaker, total 2000 utterances. Performance analysis of different feature extraction methods are compared using DTW based speech recognition method. For the purpose of evaluating the robustness against environmental additive noise, different types of noise were added to the test data with different Signal to Noise Ratios (SNR) including 0, 5, 10 and 20 dB. The typical types of noise include babble noise, car noise, airport noise, street noise, HF channel noise, exhibition noise and station noise are extracted from AURORA database. In the training phase, clean signal is used for creating template model. In the testing phase, clean signal is added with different type of noises. The noisy utterances are enhanced by Kalman filter and used for recognition. Performance of average recognition accuracy rate of noisy speech in different recognition methods is compared [12].

The performance of the recognizer was measured in terms of word accuracy (WAC):

$$WAC = \frac{N - I}{N} \times 100$$

where:

$N$ = The total no. of words in the test set

$I$ = The number of insertion errors.

Different feature extraction methods like MFCC, LPCC and BFCC are used for analysis. For those methods, the performance of average recognition accuracy rate is calculated using DTW based speech recognition method.

![Comparison of Speech Recognition System Accuracy for Different Feature Extraction Methods](image)
From the Fig.1, highest recognition accuracy is obtained for MFCC with 89% accuracy. Next high recognition accuracy rate achieved for BFCC with 88%. LPCC has lower than other feature extraction methods. From the analysis, MFCC is achieved better performance, so it is preferred for robust speech recognition system.

Performance comparison of average recognition accuracy rate of noise speech in different recognition methods is shown in Fig.2. DTW and VQ\HMM are two recognition methods are used for comparing the average recognition accuracy rate.

Table 1 shows the recognition accuracy of digits from zero to nine for airport noise using DTW modeling with different SNR values. From the average accuracy rate, digit four has highest degree of accuracy with 66.75% and three has lowest accuracy rate.

Similarly Table 2 shows recognition accuracy of digits from zero to nine for babble noise in VQ\HMM modeling with different SNR values. Digit three results lowest accuracy rate and seven has highest accuracy.

Table 3 shows the performance analysis of each noise is compared based on average recognition accuracy of two different speech modeling methods with different SNR values such 0dB, 5dB, 10dB, 15dB are considered. Noises considered for analysis are babble noise, street noise, HF channel noise, exhibition noise and station noise respectively. In DTW modeling technique exhibition noise with 15dB value has achieved highest accuracy rate and in HMM modeling technique street noise has achieved has accuracy rate.

Overall average accuracy of different noisy speech is calculated for each word and compared with different recognition methods. In VQ\HMM, high performance is achieved for digit seven with 60% accuracy. In DTW, high accuracy rate is obtained for digit one with 65% accuracy in noisy environment. The next better accuracy rate achieved for digit four with 62% accuracy. The least recognition accuracy is obtained for digit three in both methods.
VII. CONCLUSION

In this paper, the improvement in speech recognition accuracy for the utterance of digits zero to nine in English under noisy environment has been considered. The utterances are added with different types of noises at various input SNR levels. Different feature extraction methods are implemented and compared their performance based on average recognition accuracy rate. From the analysis, MFCC has obtained better recognition accuracy rate with 89%. In noise robust recognition system, testing signal is added with different types of noises and their features are extracted using MFCC. Under clean condition, DTW has achieved 89% accuracy and HMM modeling has achieved 75% accuracy rate. Performance of DTW and VQ\HMM are compared based on recognition accuracy rate in noisy environment. The average recognition accuracy obtained for DTW is 69% and VQ\HMM has obtained the accuracy rate of 65%. DTW based recognition system is somewhat better than the HMM recognizer in noisy environment. In future, different type of enhancement methods will be implemented and their accuracy will be calculated in different environments.

REFERENCES