Speech Recognition System: A Review

Nitin Washani
M. Tech Scholar
DIT University
Dehradun (Uttarakhand, India)

Sandeep Sharma, Ph.D.
Head of Department of ECE
DIT University
Dehradun (Uttarakhand, India)

ABSTRACT
To be able to control devices by voice has always intrigued mankind. Today after intense research, Speech Recognition System, have made a niche for themselves and can be seen in many walks of life. The accuracy of Speech Recognition Systems remains one of the most important research challenges e.g. noise, speaker variability, language variability, vocabulary size and domain. The design of speech recognition system requires careful attentions to the challenges such as various types of Speech Classes and Speech Representation, Speech Preprocessing stages, Feature Extraction techniques, Database and Performance evaluation. This paper presents the advances made as well as highlights the pressing problems for a speech recognition system. The paper also classifies the system into Front End and Back End for better understanding and representation of speech recognition system in each part.

General Terms
Energy, Correlation, Zero Crossing Rate, etc.

Keywords
VAD, Feature Extraction, Hidden Markov Model, Neural Networks.

1. INTRODUCTION
Since ages speech has been an important mean of communication between humans. Speech Recognition is the process of converting an acoustic speech into text, and / or identification of the speaker.

Over the years with recent advent in technology it has become an essential and integral part of our lifestyle due to the increasing communication between human and computers or automated systems [1-3].

A system built at Bell Laboratory in 1952 which was the first word recognition system which was trained to recognize digits [3]. Some of the widely used speech recognition systems are Types of Speech Recognition Systems. Some of Speaker Dependent Systems, Speaker Independent System, Isolated Word Recognizer, Connected Word Recognizer, and Spontaneous Recognition System.

Over the years the Speech Recognition Systems have come a long way the process has ensured its presence due to the well-established need of voice operated systems. However, there is a lot to be accomplished. Most of research done so far is attributed to the fact that speech is a very subjective phenomenon. The general known problems are Speaker Variation, Background Noise and Continuous Character of Speech. Perhaps the most evident source of performance degradation in speech recognition is Noise. Noise can beclassified as either environmental i.e. traffic, rain, other people talking or speaker included i.e. coughing, sneezing, swallowing, breathing, chewing, etc.

In this article, Speech Recognition System has been subdivided into Front-End and Back-End (as shown in Figure 1 below), based on the subdivision a brief review of work done so far in the domain of speech recognition system has been presented.

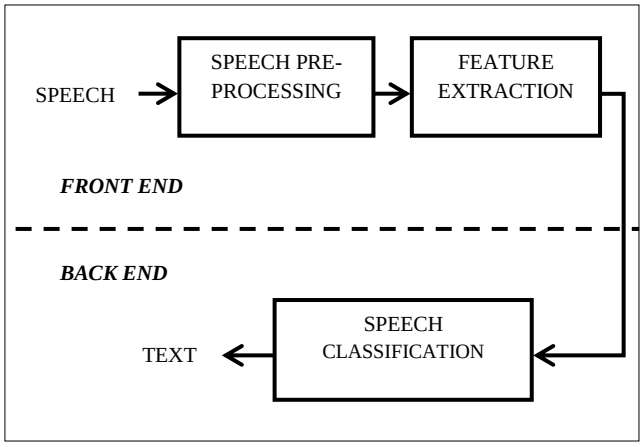

Figure 1. Speech Recognition System

2. FRONT-END ANALYSIS
Front-End of the speech recognition system comprises of Speech Preprocessing and Feature Extraction Block. Noise and differences in Amplitude of the signal can hardly influence the integrity of a word while timing variations can cause a large difference amongst samples of the same word. These issues are dealt with in the Signal Preprocessing part. Preprocessing generally involves End Point Detection, Pre-emphasis Filtering, Noise Filtering, Framing, Windowing, Echo Cancelling, etc. [4]. Block Diagram for Signal Preprocessing stage is shown in Figure 2 below.

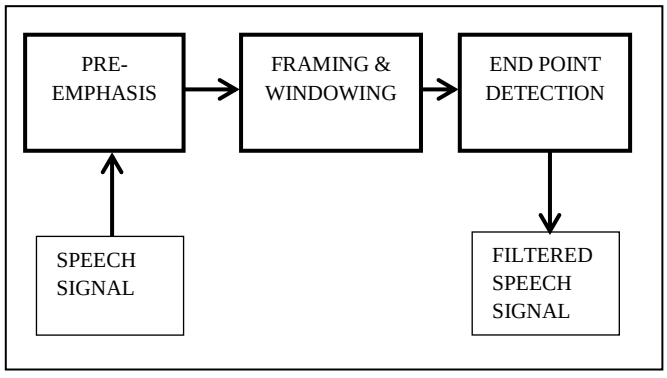

Figure 2. Signal Preprocessing

Feature Extraction is a process extracting specific features of the preprocessed speech signal. This can be done with numerous types of Techniques like Cepstrum analysis, Spectrogram, MFCC (Mel Frequency Cepstrum Coefficient), LPC (Linear Predictive Coefficient), etc. In 1976, L.R.

Rabiner, M.J. Cheng, A.E. Rosenberg and C.A. Mcgonegal gives the detailed comparative performance analysis of seven Pitch detection Algorithm [5] i.e. Modified Autocorrelation Method using clipping(AUTOC), Cepstrum method(CEP), Simplified Inverse Filtering Technique(SIFT), Data Reduction method(DARD), Parallel Processing method(PPROC), Linear Predictive Coding(LPC) and Average Magnitude Difference function(AMDF). The performance Strength and Weakness of each of the pitch detectors for different speaker w.r.t different parameters is shown in Figure (3-6).

Later in 1999, S. Ahmadi and A.S. Spanias presented an Algorithm which improves the detectability of low frequency pitch peaks by using signal dependent initial threshold and a different Cepstral weighting function [6]. Preemphasis of the speech signal brings about deterioration in the vowel recognition performance because vowels lie in the lower frequencies region and Preemphasis put undue weight on

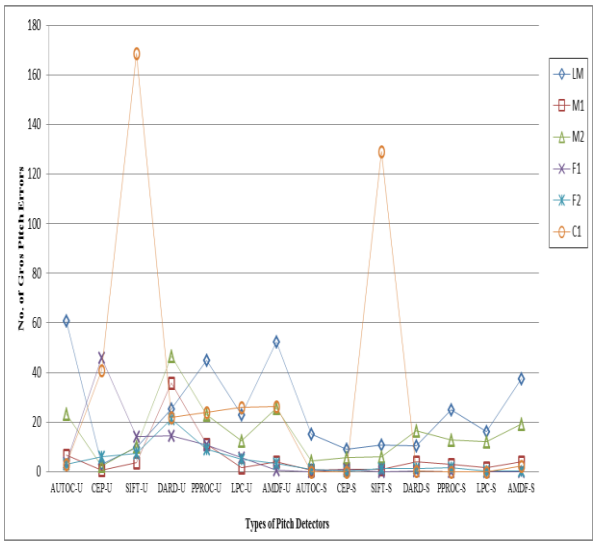

Figure 3.Number of Gross Pitch Errors-Unsmoothed/Smoothed

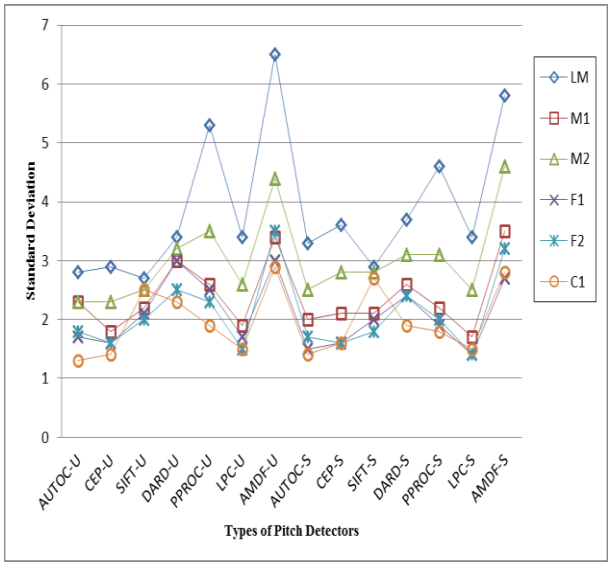

Figure 4. Standard Deviation of Fine Pitch Errors-Unsmoothed/Smoothed

In 2011, A.N. Mishra, M. Chandra, A. Biswas and S.N. Sharan performs comparative analysis of Feature Extraction methods i.e. MFCC, ΔMFCC, BFCC, PLP, RPLP and MF-

higher frequency components. In [7] we study the effect of Preemphasis for four different distance measures i.e. Euclidean, Correlation, Mahalanobis and Itakura. Without Preemphasis Itakura performs well giving 94% recognition rate and with Preemphasis Mahalanobis performs well giving 92.4%. To reduce the bad effects of the Preemphasis in the case of voiced sounds a new Algorithm is introduced in [8] which respect the fundamental concept of the Mel scale while keeping unchanged the initial frequency resolution obtained after FFT.

In 2010, I. Patel and Dr. Y.S. Rao designed an efficient Speech Recognition system by using a MFCC as a Feature Extraction with Frequency Sub-band Decomposition Technique [9].These modified MFCC performs more accurate than MFCC without Sub-band Decomposition. This system can perform more accurate by giving stress on Signal Preprocessing stage.

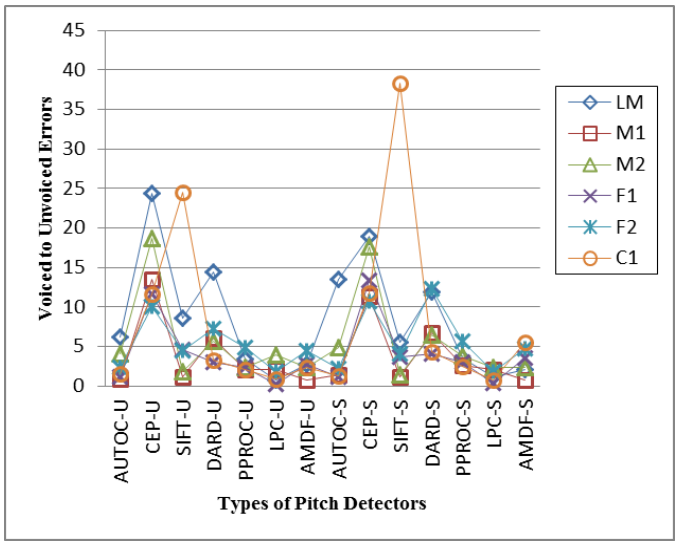

Figure 5.Voiced to Unvoiced Errors (Wideband Data)-Unsmoothed/Smoothed

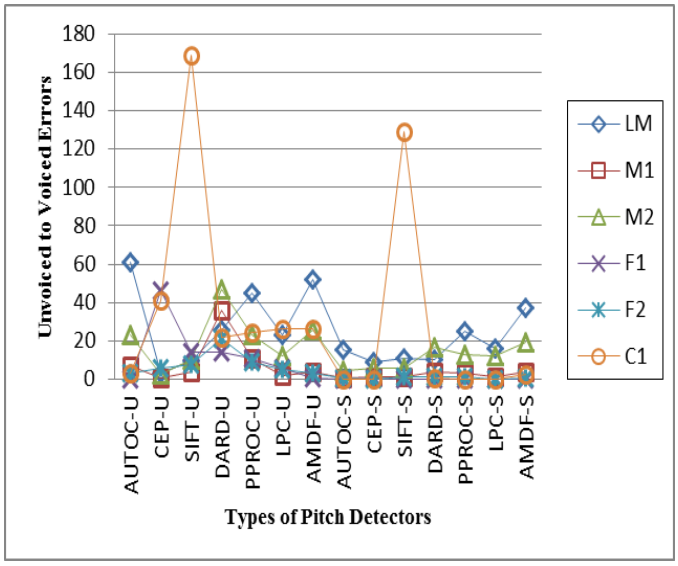

Figure 6.Unvoiced to Voiced Errors (Wideband Data)-Unsmoothed/Smoothed

PLP on the basis of Recognition accuracy for Speaker Independent Connected Hindi Digits Recognition systems [10]. This system is experimented in both Clean and Noisy

Environments. Here we can observe that MF-PLP is best and efficient Feature Extraction method for the system.HMM is used as a Classifier for which HTK is used. For different Noisy Environment different types of Noise from NOISEX-92 database has been added to clean Hindi Digits database.

In [11], S. Furui staff of the Acoustics Research Dept. at Bell Laboratories, New Jersey Designed a Speaker Independent Isolated Word Recognition with Recognition Rate of 97.6%. Here Dynamic Spectral features are used for Word Recogniton.These technique is combination of instantaneous and dynamic features of the Spectrum. Here End Point Detection is done on the basis of Autocorrelation concept. In 1999, S. Ahmadi and A.S. Spanias presented an improved method for Voiced/Unvoiced classification based on Statistical Analysis of Cepstral peak, Zero Crossing Rate and Short Time Energy of speech signal. This algorithm applied on large speech database i.e. TIMIT (under noisy condition also), in result it gives better performance compared to conventional Cepstrum method [6]. Later in 2004, R.G. Bachu, S. Kopparthi, B. Adapa and B.D. Barkana proposed simple and efficient approach for discrimination of Voiced/Unvoiced part of Speech by using Zero Crossing Rate and Energy concept together [12]. In 2012, Anand Singh explained the effects of EPD algorithm in Speech Recognition System by giving comparative analysis Time duration, Number of Samples, Root Mean Square and Mean Power with and without EPD algorithm [13]. Here 4 males and 4females Speakers uttered 50 words in 4 different Moods (i.e. Normal, Happy, Anger and Surprise).Hence a development of Database has also been described.

In [14], K. Waheed, Kim Weaver and F.M. Salam proposed a robust algorithm for Endpoint Detection using an Entropic concept. This algorithm results better than Energy based algorithm of about 25% in case of Isolated Speech and 16% in case of Connected Speech. This algorithm also provides good result in Noisy Environment. Later in [15], a new hybrid Algorithm is proposed for Isolated Word EPD by Lingyun Gu and S.A. Zahorian. This Algorithm uses a concept of Teager Energy and Energy Entropy features. Here Teager Energy is used to determine Raw Endpoints and Energy-Entropy method is used to make a final decision.This algorithm also works well in Noisy Environment. In 2002, Qi Li, J. Zheng, A. Tsai and Qiru Zhou proposed two End Point Detection Algorithms for Real Time Speech and Speaker Recognition. This article gives comparative analysis of Word Error rate in Adverse Noisy Environment [16]. These Algorithms are Reliable and Robust at various Noise levels. A low Computational complexity and fast response time are main advantage of these algorithms. Later in 2011, An Efficient algorithm for EPD of Isolated Word and Digits of different Languages by using concept of Short Time Energy and Zero Crossing Rate is proposed by Nitin N. Lokhande, Navnath S. Nene and Pratap S. Vikhe [17].This algorithm also reduces computational time and memory requirements.

In 2000, H. Jiang ,K. Hirose and Q. Huo proposed a Novel approach to perform Quasi-Minimax decision rule in Continuous Speech Recognition i.e. Minimax Recursive Search Algorithm [18]. In this Article there is comparison of Viterbi Algorithm and Minimax Algorithm on the basis of Recognition accuracy and Computational complexity in present of Noise. Later in [19], J.K. Lee and Chang D. Yoo proposed Speech Enhancement technique based on Wavelet Transform (WT).The main objective of speechEnhancement is to reduce Noise by minimizing speech distortion. In the proposed Speech Enhancement algorithm different threshold values are used for Voiced and Unvoiced frames, which results in better performance as compared to traditional Speech Enhancement methods.

3. BACK-END ANALYSIS

Back-End consists of Speech Classification block. Speech Classification process is for classifying the extracted features and relates the input sound to the best fitting sound from a database and represents them as an output. The commonly used techniques for Speech Classification are HMM (Hidden Markov Model), DTW (Dynamic Time Warping), VQ (Vector Quantization), ANN (Artificial Neural Network), etc. [3]. In many Speech recognition systems, hybrid techniques are implemented and work in a cooperative relationship. Neural Networks perform very well at learning phoneme probability from highly parallel audio input, while Markov Models can use the phoneme observation probabilities that Neural Networks provide to produce the likeliest phoneme sequence or word.

In [20], Comparison of two different types of Neural Networks i.e. Multi-Layer Feed Forward and Radial Basis Function Network for Speech Recognition when Mel-Frequency Cepstrum Coefficient is used in Signal Preprocessing stage. Here RBF network needs more amount of Hidden Layer as compared to Multi-Layer Feed Forward Network and increase in Number of Hidden Layer increases Computational time of system. In 2014, Amr Rashed gives comparative analysis of different Neural Network Learning Algorithms [21]. A Feed Forward Multi-Layer Perceptron Neural Network algorithm gives fast and accurate result even in presence of Whit Gaussian Noise. Here we can also observe that Sequential Weight/Bias training algorithm gives efficient result in Speech Recognition Systems.

T. Lee, P.C. Ching and Lai-Wan Chan propose a novel approach of utilizing Recurrent Neural Network (RNN) for Isolated Word Recognition [22]. Here the RNN Speech Model is trained in two stages. First, The RSM‟s are trained independently to extract the Temporal and Static characteristic of individual words. Second, Mutual discriminative training among the RSM‟s takes place for minimizing the probability of misclassification and improving the recognition accuracy. In 2012, K. Dutta and K.K. Sarma proposed a combined architecture of LPC and MFCC Feature Extraction technique by two different RNN. This combined Architecture gives 10% gain in recognition rate than individual architectures. Hence the result obtained by the proposed system can be easily improved by using Hidden Layered based RNN [23]. In Parallel they have also proposed a Dynamic Segmentation of Voiced/Unvoiced segments from speech utterance [24]. This results in 90% Recognition Rate but the Testing Time of the system is increased by small amount which can be counterbalance byusing parallel processing technique results in improvement of speed of computation.

In 2012, Anand Singh, D.K. Rajoriya and Vikas Singh uses LPCC as Feature Extraction method and ANN as Classifier for Speech Recognition of Hindi Hybrid Paired words and observes that Consonant dominated words provides better Recognition rate as compared to Vowel dominated words [25]. In [4], Comparative analysis of different training algorithm is presented in which “trainscg” training algorithm performs well over a wide variety of problems. Here for efficient Speech Recognition System Neural Network with MFCC is used. The result can be improved by increasing the training data size. In 2013, Manan Vyas designed a GMM based Speaker-Dependent Speech Recognition System [26]. In this System End Point of speech utterance is detected by

concept of Energy & ZCR and MFCC is used as feature extraction technique. Hence GMM gives a poor Recognition Rate (70%) as compared to other classifiers, but in case of Speaker Recognition it gives efficient results.

4. CONCLUSION

Speech Recognition is a challenging problem to deal with. We have attempted in this paper to provide a review of how much this technology has progressed in the previous years. The performance of Speech Recognition System is mainly depends on the quality of Signal Preprocessing Stage. The Preprocessing quality is giving the biggest impact on the Speech Classification performance. Signal Preprocessing consist an EPD, Filtering, Framing, Windowing, Echo Cancellation, etc. An Improvement in any individual part can improve the overall system performance. For effective working of Back-End there should be more efforts in Front-End processing. MFCC is more preferred in Feature Extraction technique as it generates the training vectors by transforming speech signal into frequency domain, and therefore it is less affected by noise.

5. REFERENCES

[1] Dr. Shaila D.Apte, "Speech and Audio Processing",Wiley India Edition.

[2] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang, "Springer Handbook of Speech Processing", Springer.

[3] L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice Hall Signal Processing Series.

[4] N. Srivastava, "Speech Recognition using Artificial Neural Network", IJESIT, Volume 3, Issue 3, May 2014.

[5] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-24,No. 5, October 1976.

[6] S. Ahmadi and A. S. Spanias, "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm", IEEE Transactions on Speech And Audio Processing, Vol. 7, No. 3, May 1999.

[7] K.K. Paliwal, "Effect of Preemphasis on Vowel Recognition Performance", Elsevier Science Publishers B.V. (North-Holland), Vol. 3. No. 1. April 1984.

[8] R. Vergin, Douglas O'Shaughnessy and A. Farhat, "Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 7, No. 5, September 1999.

[9] I. Patel, Dr. Y. Srinivas Rao, "Speech Recognition Using HMM with MFCC-AN Analysis Using Frequency Spectral Decomposition Technique", SIPIJ,Vol. 1,No. 2,December 2010.

[10] A. N. Mishra, M. Chandra, A. Biswas, S. N. Sharana, "Robust Features for Connected Hindi Digits Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 2, June, 2011.

[11] Sadaoki Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSp-34, No. 1, February 1986.

[12] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D., "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal", Springer Science & Business Media.

[13] A. Singh, Dr. D. K. Rajoria, V. Singh, "Database Development and Analysis of Spoken Hybrid Words Using Endpoint Detection", IJECSE, Volume 1, Number 3.

[14] K. Waheed, Kim Weaver and F. M. Salam, "A Robust Algorithm for Detecting Speech Segments Using an Entropic Contrast".

[15] Lingyun Gu and S. A. Zahorian, "A New Robust Algorithm for Isolated Word Endpoint Detection".

[16] Qi Li, J. Zheng, A. Tsai and Q. Zhou, Member, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition", IEEE Transactions on Speech And Audio Processing, Vol. 10, No. 3, March 2002.

[17] N. N. Lokhande, N. S. Nehe, P. S. Vikhe, "Voice Activity Detection Algorithm for Speech Recognition Applications", ICCIA, 2011.

[18] Hui Jiang, K. Hirose and Qiang Huo, "A Minimax Search Algorithm for Robust Continuous Speech Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 8, No. 6, November 2000.

[19] J. K. Lee and C. D. Yoo, "Wavelet Speech Enhancement Based On Voiced/Unvoiced Decision", the 32nd International Congress and Exposition on Noise Control Engineering Jeju International Convention Center, Seogwipo, Korea, August 25-28, 2003.

[20] W. Gevaert, G. Tsenov and V. Mladenov, "Neural Networks used for Speech Recognition", Journal Of Automatic Control, University Of Belgrade, Vol. 20:1-7, 2010.

[21] Amr Rashed, "Fast Algorithm for Noisy Speaker Recognition Using ANN", IJCET, Volume 5, Issue 2, February (2014), pp. 56-65.

[22] T. Lee, C. Ching and Lai-Wan Chan, "Isolated Word Recognition Using Modular Recurrent Neural Networks", Pattern Recognition, Vol. 31, No. 6, pp. 751—760, 1998.

[23] K. Dutta and K. K. Sarma, "Multiple Feature Extraction for RNN-based Assamese Speech Recognition for Speech to Text Conversion Application", International Conference on Communications, Devices and Intelligent Systems (CODIS), IEEE, 2012.

[24] K. Dutta and K. K. Sarma, "Dynamic Segmentation of Vocal Extract for Assamese Speech to Text Conversion using RNN", CISP, IEEE, 2012.

[25] A. Singh, Dr. D. K. Rajoria, V. Singh, "Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks", International Journal of Computer Applications, Volume 52, No.12, August 2012.

[26] M. Vyas, "A Gaussian Mixture Model Based Speech Recognition System Using Matlab", SIPIJ, Vol.4, No.4, August 2013.

[27] Hiroaki Sakoe, "Two-Level DP-Matching, A Dynamic Programming Based Pattern Matching Algorithm For Connected Word Recognition", IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-27, No. 6, December 1979.