Speech recognition in noisy environment, issues and challenges: A review

Conference Paper · February 2015

DOI: 10.1109/ICSNS.2015.7292420

Ujwalla Gawande
Yeshwantrao Chavan College of Engineering
28 PUBLICATIONS **210** CITATIONS

2015 International Conference on Soft-Computing and Network Security (*ICSNS* -2015), Feb. 25 – 27, 2015, Coimbatore, INDIA

Speech Recognition in Noisy Environment, Issues and Challenges: A Review

Karishma Chavan
Department of Computer Technology
YCCE Nagpur, India
karishma.chavhan78@gmail.com

Ujwalla Gawande
Department of Computer Technology
YCCE Nagpur, India
ujwallgawande@yahoo.co.in

Abstract—**The voice is most prominent & primary mode of communication among the human beings. With this speech human can communicate with machine, thus this technique is used in education, military and medical sectors. Though this is not the new area, from last few decades researchers are working on the improvement of accuracy in voice recognition system. The design of that system concerns major issues such as speech classes, speech styles, vocabulary, transducers, illness and channels; due to all this constraints the noise factor in automatic speech recognition is high. Many researchers have tried to overcome above issues. The recent work in speech recognition is briefly summarized in this paper. Some basic fundamentals of clean and noisy databases of voice samples are also discussed.**

Keywords—Noisy and Clean database;feature extraction; feature recognition.

I. Introduction

Speech recognition is the task to identify spoken words and convert it into machine readable and understandable format.The advances in digital signal processing technology are used in many different application areas of speech processing like speech signal compression, enhancement, synthesis, and recognition [1, 2]. Speech recognition plays important role in many applications [4] which is depicted as follows.

Education purpose: For teaching students of foreign languages to pronounce vocabulary correctly.

***Medical sector*:**For health care.

***Military sector*:**For high performance fighter aircraft, helicopters, battle management.

***Translation*:**Fortranslation from one language to another.

Though this technology has been well developed technique, still there are some issues which make a system less accurate which are discussed below.

Environment: Due to type of noise, signal/noise ratio, working conditions and weather is changing day by day so our voice also changes according to weather conditions.

Transducer: Due to microphone, telephone like changing in frequency range of caller and receiver.

Channel: Due to band amplitude, distortionechoes.

Vocabulary: Due to characteristics of available training data, specific or generic vocabulary.

Speech styles: Due to voice tone like quiet, normal, shouted and speed of voicelike slow, normal, and fast

Illness: Due tocough, fever.

The main reason of the above mentioned issues are the noisy environment. Therefore the main objective of current speech recognition system is to develop accurate voice recognition system, which will work on the above issues.This system can work better in noisy environment.

The paper is organized as follows, section1consist of introductionary part, in section 2 we describe a brief survey of Automatic Speech Recognition (ASR). A related technique ofASR discussed in section3.Section 4 contains the performance analysis. Finally conclusion is in section 5.

II. LITRATURE REVIEW

Speech is the most natural way of human communication and its processing has been one of the most exciting research areas of signal processing. Though there are major advances in statistical modelling of speech, Automatic Speech Recognition (ASR) today has wide applications that require human machine interface. In this section a brief review of major highlights during last seven decades in the research and development of ASR is discussed.

The first speech recognition has been started by recognizing the digits only [1]. The spoken digits were recognizing by Audrey system in 1952 at bell laboratories.IBM introduces Shoe Box in 1962, which recognizes 16 English worlds. Speech recognition has been expanding day by day and to recognize about a few hundred words. That has the potential to recognize an unlimited number of words because of new statistical method known as the Hidden Markov Model (HMM) [1].In1990’s the methods for statistical learning of acoustic, language models forstochastic language understanding and the methods for implementation of large vocabulary speech understanding systems was introduces.After that, the computer speech recognition systems was introduced and performed well. To overcome these issues within arbitrary environment they still had a problem with the pitch level i.e. low, high, among similar-voice sample words. In 2008 the key technologies developed have to recognize very large vocabulary for limited task within arbitrary environment. Recentresearchers are working onunlimited vocabulary using ASR for unlimited task and for many languages.SantoshK.Gaikwad, et al. [2] discussed the brief overview of all speech recognition techniques. An efficient algorithm for extracting speech was proposed byWei HAN et

978-1-4799-1753-2/15/$31.00 ©2015 IEEE

al.[3], which gives computation power 53% and accuracy claimed was 92.93%, where FFT filter has been used for enhancement. M.A.Anusuya et al. [4] discussed about all speech recognition techniques and gives better future scopes for implementation of speech recognition system.Mohammad A. M. Abushariah et al. [5], usedHidden Markov Model (HMM) asclassifier. MFCC has been used for feature extraction byusing English digits database, with recognition rate up to92%.Hui Jiang et al.[6] proposed a method for estimating continues density HMM for ASR by the principle of maximizing the minimum multiclass separation margin and gave significant recognition error rate. Spectral subtraction approach and signal subspace approach had drawback of residual noise therefore, M. A. Abd El-Fattah etal. [7], used adaptive wiener filtering approach which gave better accuracy.Leena R Mehta et al. [8] compared the Mel Frequency Cepstral Coefficients(MFCC) and Linear Predictive coding(LPC) methods .They foundMFCC is better than LPC.

The techniques used in speech recognition system are discussed in section 3 according to their working nature.

III. RESEARCH METHODOLOGY

There are four main steps used for speech recognition shown in Fig1. First step is the creation of database, which contain voice samples those will be recorded in quiet and noisy environment. Second, Speech pre-processing (enhancement and normalization), where unwanted noise is removed. Third, feature extraction and last is matching, which is based on percentage of the speech being recognized.

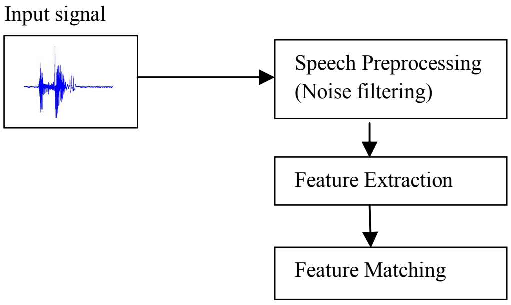

Fig 1: Block diagram of Speech Recognition System

Each of these steps discussed below.

A. Voice sample Database

In speech recognition system standard database are available, but every author create their own voice sample database according to their requirement of problem definition. At the time of voice recording some key points about voice frequency should be considered.A voice frequency or voice band is one of the frequencies, within part of the audiosignal. Voice frequency band ranges from approximately 300 Hz to 3400 Hz. The bandwidth of single voice-frequency transmission channel is usually 4 kHz that including guard bands, and samplingfrequency is 8 kHz. The fundamental frequency of adult male is 85 to 180 Hz, and adult female is 165 to 255 Hz.

The system should collects speech samples of different speakers.The samples should be from male or femalebelonging to different ages. These samplesshould be recorded in clean environment by using specialized Visual- Audio studio. For real environment some samples can be recorded in noisy environment, like faculty class rooms and students' rooms in the university's hostels.

B. Speech Pre-processing

Pre-processing is the task to remove noise from speech signal.Speech is generally a random, so for the same speaker; the same words may have different frequency bands. This is due to the different vibrations in vocal cords. Thus, the shapes of frequency spectrum generated may be different. But, the similarity between these spectrums determines the degree of recognition between the speech signals.

There were many speech pre-processing techniques available to improve the recognition performances. Speech enhancement is one of the most important topics in speech signal processing. Several techniques have been proposed for this purpose like the spectral subtraction approach [7], the signal subspace approach [9] and adaptive wiener filtering approach [10]. The performances of these techniques depend on quality of the processed speech signal. The improvement of the speech Signal-to- Noise (SNR) ratio is the target of most techniques.

Spectral subtraction approach is the earliest method for enhancing speech degraded by additive noise [7]. This technique estimates the spectrum of the clean (noise-free) signal by the subtraction of the estimated noise magnitude spectrum from the noisy signal, while keeping the phase spectrum of the noisysignal. Residual noise is the drawback of this technique.

Another technique is a signal subspace approach [9]. This is used for enhancing a speech signal degraded by uncorrelated additive noise or colored noise. The idea of this algorithm is based on the fact that the vector space of the noisy signal can be decomposed into a signal plus noise subspace and an orthogonal noise subspace. Processing is performed on the vectors in the signal plus noise subspace only, while the noise subspace is firstly removed. Decomposition of the vector space of the noisy signal is performed by applying an eigenvalue or singular value decomposition or by applying the Karhunen-Loeve transforms (KLT). A. Rezayee et. al. [11]has proposed the signal / noise KLT based approach for colored noise removal. Idea of this approach is that noisy speech frames were classified into speech-dominated frames and noise-dominated frames. The signal KLT matrix has been used in the speech-dominated frames andthe noise KLT matrix has been used in the noise-dominated frames.

The wiener filter is a popular technique that has been used in many signal enhancement methods. The drawback of the Wiener filter is the fixed frequency response at all frequencies and the requirement to estimate the power spectral density of the clean signal and noise prior to filtering. Therefore M. A. Abd El-Fattah et al.[10],proposedan adaptive Wiener filter approach for speech preprocessing.This approach depends on the adaptation of the filter transfer function from sample to sample based on the speech signal statistics (mean and variance). This approach provides the best SNR improvement among the spectral subtraction approach and the traditional Wiener filter approach in frequency domain. This approach can also treat musical noise better than the spectral subtraction approach and it can avoid the drawbacks of Wiener filter in frequency domain.

C. Feature extraction

In speech recognition, the main goal of the feature extraction step is to compute a mean sequence of feature vectors which provide a compact representation of the given input signal.Previously the various techniques were used for feature extraction invoice recognition such as Linear discriminant analysis (LDA)**,** Linear Predictive coding (LPC), Mel Frequency Cepstral Coefficients (MFCC).Thesetechniques arebriefly discussed next.

1) ***Linear discriminant analysis (LDA):***LinearDiscriminant Analysis (LDA) is originally used for classification. The LDA has been used to improved recognition performance for small-vocabulary or considerable large vocabulary [12].

The LDA consider two feature vector X and Y. Find a linear transformation of feature vectors X from an n-dimensional space to vectors Y in an m-dimensional space (m<n)[9]. The class seperabiliy of this feature vector is large therefore scatter matrices are used to formulate the optimization problem. Consider two matrices S1, S2 out of the three - W: within-class, B: between-class, T: total scatter matrix,are used where many combinations are possible [20]. Several optimization criteria are also possible, the most widely used ones are to maximize

$$J1(m) = tr(S_{2y}^{-1}S_{1y}) \quad (1)$$

$$J2(m) = \det\,(S_{2y}^{-1}S_{1y})(2)$$

Wheretr $(S_{2y}^{-1}S_{1y})$denotes the trace of A, det $(S_{2y}^{-1}S_{1y})$ its determinant and Siyis the scatter matrix in the mdimensional y-space. They chose S1= T, S2 = W, so that's why they consider a class-independent linear transformation ofthe vector space. That enhances the total scatter while keeping within-class scatter constant. The optimization of (1) and (2) leads to the result .That the input vector **x** has to be projected onto the subspace spanned by those m eigenvectors of$S_{2x}^{-1}\,S_{1x}$, which correspond to the m largest eigenvalues. Note that J1 and J2 lead to the same set of features. Those criteria are invariant under any non-singular linear transformation both in the original n-dimensionalspace and in the resulting m-dimensional space. Even the resulting features are the same irrespective of any linear transformation in the n-dimensional space prior to the LDA transformation. In this way the LDA feature extraction method worked. The main drawback of this method is that, not give clear voice for large vocabulary. Therefore current researchers think about LPC and MFCC methods.

2) ***Linear Predictive coding (LPC):***Linear Predictive coding (LPC) is a computational mathematical operation that analyses the speech signal by estimating the formants which remove their effects from the speech signal and estimate thefrequency and intensity. In LPC, each sample of the signal is expressed as a linear combination of the previous signal frames. This is called a linear predictor and hence it is called as linear predictive coding .Fig2 shows the steps of LPC.

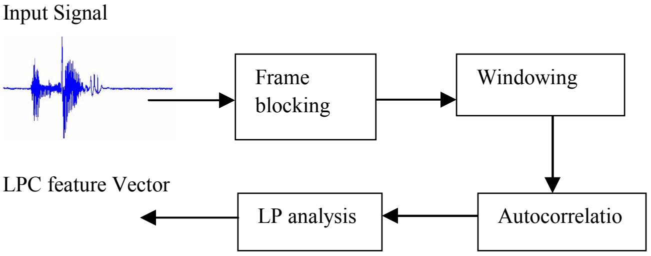

Fig2: Process Linear Predictive coding

In [8], Leena R. Mehata proposedLPC for feature extraction and design of a speaker-independent speech recognition system for Marathi Language. LPC has four steps firstly divide speech signal into frames, overlap its previous step by predefined size .The second step is to window all frames, this is done in order to eliminate discontinuities at the edges of the frames. Next step is to take Fast Fourier Transform with autocorrelation of each frame. This technique is a fast way of Discrete Fourier Transform and it changes the domain from time to frequency. In this way the LPC gives better recognition rate nearly up to 90%.

3) ***Mel Frequency Cepstral Coefficients (MFCC):***Mel Frequency Cepstral Coefficient (MFCC) s is one of the most accurate feature extraction method used in automatic speech recognition. Feature vectors are extracted from the frequency spectra of the windowed speech frames Therefore technique is called FFT based feature extraction method. The Mel frequency filter bank is a series of triangular band pass filters[8]. The filter bank containing a non-linear frequency scale called the mel-scale.Fig 3 shows the steps of MFCC.

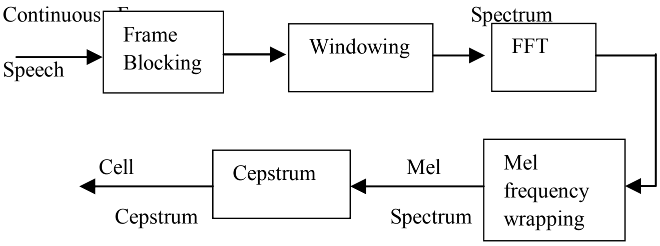

Fig 3: Process Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) Processing:

Frame Blocking:In first step the signal is partitionedor blocked into N segments we called it frames. Framing is the first applied to the speech signal of the speaker.

Windowing:Inthe second step in the processing is to window each individual frame, so as to minimize the signal discontinuities at the beginning and end of each frame.

Fast Fourier Transform:Next step is the Fast Fourier Transform which converts each frame of N samples in time domain to frequency domain.

Mel-Frequency Wrapping:The spectrum obtained from the above step is Mel Frequency Wrapped; the major work done in this process is to convert the frequency spectrum to Mel spectrum.

Cepstrum:In final step, convert the log Mel spectrum back to time. That result is called the Mel frequency Cepstrum coefficients (MFCC).

In [8], Leena R. Mehta compared the performances of MFCC and LPC features under VQ environment. Finding nearby 78%, 85% accuracy respectively. A result lead to MFCC is better than LPC.

D. Feature recognition technique

The goal of feature recognition is to divide interested object into one of a number of categories. The object of interest known as pattern*s* and the classes refer to the individual words. Thus the procedure appliedon the extracted features is called as Feature matching or pattern matching. There were two methods used for Feature recognition,are briefly discussed in this paper.

1) ***Hidden Markov Model (HMM):*** HMM is more famous because they can be trained automatically and are simple and computationally feasible to use. GhulamMuhammadet al. [17] had used the HMM to find out handwritten wordsextracted from a tablet. HMM's be compatible with to complete words can be easily developed (with the help of pronunciation dictionary) from phone HMM's and word sequence probabilities added. Complete network searched for best path corresponding to the optimal word sequence. HMMs are very simple networks technique. That can produce speech by using a number of states for each model and modelling the short-term spectra associated with each state with, generally, mixtures of multivariate Gaussian distributions (the state output separation)[6]. Parameters of the model are the state transition probabilities and the means, variances and mixture weights which characterize the state output distributions. Each phoneme or word, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes.

2) ***Vector Quantization (VQ):***The most successful text independent recognition method is based on VQ environment,which is classified into two parts such as features training and featuresmatching. Features training are generally referred with randomly choosing feature vectors and perform training for the codebook using vector quantization algorithm[8]. VQ codebook has of a small number of representative feature vectors and used as an efficient means of specifying speaker-specific features. After that speaker-specific codebook is produced by grouping the training feature vectors of each single speaker. The recognition stage, an input utterance is vector-quantized by the codebook of each reference speaker. VQ distortion accumulated over the entire input utterance is used to make the recognition decision. This method is stronger than a continuous HMM method.

IV. PERFORMANCE COMPARISION OF EXISTING SYSTEMS

The performance of speech recognition systems is normally referred in terms of accuracy and speed. Usually accuracy is considered as word error rate (WER), whereas speed is considered as the real time factor. Alternately accuracy can take Single Word Error Rate (SWER) and Command Success Rate (CSR).Word Error Rate (WER), is a common measurement of the performance of speech recognition. Normally problems occurred in the performance of system due to mismatch of recognisedword sequence with reference word sequence. Operating at the word level. Word error rate calculated as

$$W.E.R = S + D + I/N$$

Where,

- S is the number of substitutions,
- D is the number of the deletions,
- I is the number of the insertions,

N is the number of words.

Sometimes word recognition rate (WRR) was Used

$$WRR = 1 - WER = N - S - D - I/N = H - I/N$$

Where,

- H is N-(S+D), the number of correctly recognized words.

TABLE1: COMPARISON OF RECOGNITION RATES.

Name of Author	Feature Extraction Technique	Feature Classification Technique	Recognition Rate
B. Milner	MFCC	VQ	88.88%
S.K.Podder	LPC	VQ and HMM	62% to 96%
S.M. Ahadi	MFCC (Clean) MFCC(Noisy)	HMM (Clean) HMM (Noisy)	86% 28% to 78%

The above table discussed the accuracy of most popular existing techniques. In [14] B.Milner discussed MFCC with VQ environment gives 88.88% recognition rate. In [15] S.K.Podder discussed about LPC with VQ and HMM environment, resulted 62% to 96% recognition rate. In [16] S.M.Ahadidiscussed MFCC with HMM in both clean and noisy environment and gives 86%, 28 to 78% recognition rate respectively. Here we found that MFCC with HMM environment is better as compared to other techniques.

V. CONCLUSION

In this review paperdifferent techniques of speech recognition are discussed. Performance of the ASR system based on thefeature extraction technique and their accuracy is compared. In coming years, the large vocabulary speaker independent continuous speech has gained more importance. Based on this review, the advantage of MFCC features is more suitable which reduces the complexity of the calculation and offersgood recognition result. It also achievereduction in time consumption.

REFERENCES

[1] Reddy, D.Raj. “Speech Recognition by Machine: A Review”Proceedings of the IEEE, vol. 64, no. 4, pp:501-531, April 1976.

[2] Santosh Gaikwad, Bharti Gawali, PravinYannawar, “A Review on Speech Recognition Technique”, International Journal of Computer Applications, vol. 10, no.3, pp”16-24, Aurangabad, November 2010.

[3] Wei HAN, Cheong-Fat CHAN, Chiu-Sing CHOY and Kong- Pang PUN, “An Efficient MFCC Extraction Method in Speech Recognition”, In Circuits and Systems, (ISCAS) Proceedings. IEEE International Symposium on pp: 4, May 2006.

[4] M.A.Anusuya, S.K.Katti,”Speech Recognition by Machine: A Review”International Journal of Computer Science and Information Security, vol. 6, no. 3, 2009.

[5] Mohammad A. M. Abushariah, ”English Digits Speech Recognition System Based on Hidden Markov Models”, IEEE International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia,11-13May 2010.

[6] Hui Jiang, Xinwei Li, and Chaojun Liu, “Large Margin Hidden Markov Models for Speech Recognition”, IEEE transactions on Audio, speech, and language processing, vol. 14, no. 5, September, pp:1584-1595, 2006.

[7] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. Acoust., Speech, Signal Processing,vol 27,no 2,pp. 113-120, 1979.

[8] Leena R Mehta, S.P.Mahajan, Amol S. Dabhade, “Comparative study of MFCC and LPC for Marathi Isolated Word Recognition system”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering ,vol. 2, no. 6,pp:2133-2139, June 2013.

[9] Y. Ephriam and H. L. Van Trees,” A signal subspace approach for speech enhancement”, in Proc. International Conference on Acoustic, Speech and Signal Processing, vol.2, pp. 355-358, Detroit, MI, U.S.A, May 1993.

[10] M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diab and F. E. Abd El-samie, “Adaptive wiener Filtering Approach for speech Enhancement” ,Ubiquitous Computing and Communication Journal, vol 3,no 2.pp;23-31,2010.

[11] A. Rezayee and S. Gazor,” An adaptive KLT approach for speech enhancement”, IEEE Trans. Speech Audio Processing, vol. 9, pp. 87-95 February. 2001.

[12] R. Haeb-Umbach, H. Ney “Linear discriminant analysis for improved large vocabulary continuous speech recognition”Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on. vol. 1. IEEE, 1992.

[13] Ujwalla Gawande, ”An efficient iris recognition system based on Efficient Multialgorothmic Fusion technique”, IJCA proceeding on international conference and workshop of emerging tread in technology (ICWET), no 13, 2011.

[14] B. Milner, “A Comparison of Front-End Configurations for Robust Speech Recognition”. ICASSP, vol. 1, IEEE 2002.

[15] S.K.Podder, “Segment-based Stochastic Modelings for Speech Recognition”, PhD Thesis. Department of Electrical and Electronic Engineering, Ehime University, Japan, 1997.

[16] S.M., Ahadi, H., Sheikhzadeh, R.L., Brennan, and G.H., Freeman, “AnEfficient Front-End for Automatic Speech Recognition”. IEEEInternational Conference on Electronics, Circuits and Systems(ICECS2003), Sharjah, United Arab Emirates, 2003.

[17] Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda , “Automatic Speech Recognition for Bangia Digits”, Proceedings of 12th International Conference on Computer and Information Technology,IEEE(ICCIT), Dhaka, Bangladesh, pp:21-23 December, 2009.

[18] Haitian Xu, Member, IEEE, Paul Dalsgaard, Zheng-Hua Tan, “Noise Condition-Dependent Training Based on Noise Classification and SNR Estimation”,IEEE transactions on audio, speech, and language processing, vol. 15, no. 8,pp:2431:2443, November 2007.

[19] Brian Kan-Wing Mak, Yik-Cheung Tam, and Peter Qi Li,”Discriminative Auditory-Based Features for Robust Speech Recognition” IEEE transactions on speech and audio processing, vol. 12, no. 1,pp:27-36, January 2004.

[20] K. Fukunaga, Introduction to statistical pattern recognition, 2nd ed., Academic Press, 1990.