A Survey on Feature Extraction and Classification Techniques for Speech Recognition

Sanjay A. Valaki, Harikrishna B. Jethva

Computer Engineering Department, L.D. College of Engineering, Ahmedabad
Sanjay.valaki@gmail.com

ABSTRACT

Speech processing is really essential research spot where speaker recognition, speech synthesis, speech codec, speech noise reduction are some of the research areas. Many of the languages have different speaking styles called accents or dialects. Speech recognition system is a way for the interface of human to machine. Automatic Speech Recognition is advance way to operate computer without much efforts through speech only. In this paper we have discuss LPC and MFCC Techniques for future extraction and some classification methods to classify after recognition of speech word like HMM and ANN. This paper is concludes with the decision on feature direction for developing technique in human computer interface system using Gujarati Language.

Keywords— *Automatic Speech Recognition (ASR), Feature Extraction, LPC, MFCCs, HMM, ANN, Gujarati Language*

I. INTRODUCTION

Human beings find it easier to communicate and express their ideas via speech. In fact, using speech as a means of controlling one's surroundings has always been an intriguing concept. For this reason, automatic speech recognition (ASR) has always been a renowned area of research. Over the past decades, a lot of research has been carried out in order to create the ideal system which is able to understand continuous speech in real-time, from different speakers and in any environment. However, the present .ASR systems are still far from reaching this ultimate goal. [1]

Over the past years, several review papers were published, in which the ASR task was examined from various perspectives. A recent review [2] discussed some of the ASR challenges and also presented a brief overview on a number of well-known approaches. The authors considered two feature extraction techniques: the linear predictive coding coefficient (LPCC) and the Mel frequency Cepstral Coefficient (MFCC), as well as five different classification methods: template-based approaches, knowledge-based approaches, artificial neural networks (ANNs), dynamic time warping (DTW) and hidden Markov models (HMMs). Finally, a number of ASR systems were compared, based on the feature extraction and classification techniques used.

A. Type of Speech

Speech recognition system can be separated in different classes by describing what type of utterances they can recognize.

1) Isolated Word

Isolated word recognizes attain usually require each utterance to have quiet on both side of sample windows. It accepts single words or single utterances at a time .This is having "Listen and Non Listen state". Isolated utterance might be better name of this class [5].

2) Connected Word

Connected word system are similar to isolated words but allow separate utterance to be "run together minimum pause between them.

3) Continuous speech

Continuous speech recognizers allows user to speak almost naturally, while the computer determine the content. Recognizer with continues speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries.

4) Spontaneous speech

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed .an ASR System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together.

B. ASR System classification

Speech Recognition is a special case of pattern recognition. There are two phase in supervised pattern recognition, viz., Training and Testing. The process of extraction of features relevant for classification is common in both phases. During the training phase, the parameters of the classification model are estimated using a large number of class examples (Training Data) During the testing or recognition phase, the feature of test pattern (test speech data) is matched with the trained model of each and every class. The test pattern is declared to belong to that whose model matches the test pattern best.

II. SPEECH RECOGNITION TECHNIQUES

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories, Davis, Biddulph and Balashek developed an isolated digit Recognition system for a single speaker [1]. The goal of automatic speaker recognition is to analyze, extract characterize and recognize information about the speaker identity. The speaker recognition system may be viewed as working in a four stages:

1. Analysis
2. Feature extraction
3. Modeling
4. Testing

A. Speech analysis technique

Speech data contain different type of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The information about the behavior feature also embedded in signal and that can be used for speaker recognition. The speech analysis stage deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting [2]. The speech analysis technique done with following three techniques

1) Segmentation analysis

In this case speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information. Study made in used segmented analysis to extract vocal tract information of speaker recognition.

2) Sub segmental analysis

Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmental analysis. This technique is used to mainly analyze and extract the characteristic of the excitation state.

3) Supra segmental analysis

In this case, speech is analyzed using the frame size this technique is technique is used mainly to analyze and characteristic due to behavior character of the speaker.

4) Performance of System

The performance of speaker recognition system depends on the technique employed in the various stages of speaker recognition system. The state of art of speaker recognition system mainly used segmental analysis, Mel frequency Spectral coefficients (MFFCs), Gaussian mixture model (GMM) and feature extraction, modeling and testing stage. There are practical issues in the speaker recognition field other technique may also have to be used for resulting a good speaker recognition performance

B. Speech Feature Extraction Technique

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances. These characteristics can be extracted

from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:

- Easy to measure extracted speech features
- It should not be susceptible to mimicry
- It should show little fluctuation from one speaking environment to another
- It should be stable over time
- It should occur frequently and naturally in speech

The most widely used feature extraction techniques are explained below.

1) Linear Predictive Coding (LPC)

One of the most powerful signal analysis techniques is the method of linear prediction. LPC [4] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech [5]. The analysis provides the capability for computing the linear prediction model of speech over time. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. The following figure 1 shows the steps involved in LPC feature extraction and Table 1 state advantages and disadvantages.

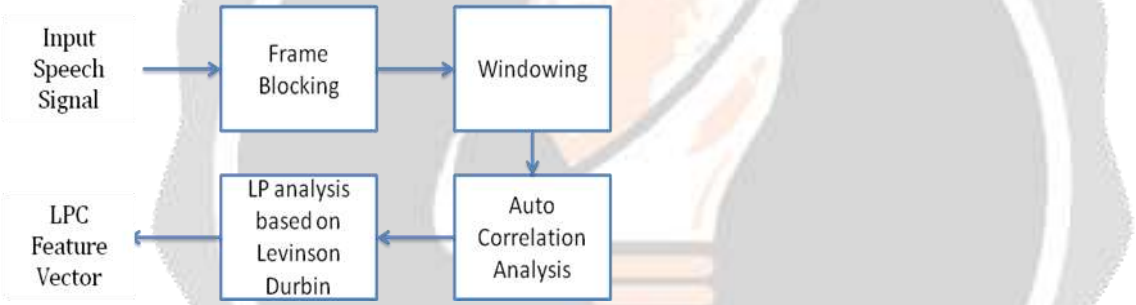

Fig. 1 Steps involved in LPC feature extraction

2) Mel Frequency Cepstral Coefficients (MFCC)

The MFCC [4] is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC [8], it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation [3][4]. It transforms the log of the quefrency domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula (1)

$$Mel(f)= 2595*\log10(1+f/700) \quad (1)$$

The following figure 2 shows the steps involved in MFCC feature extraction and table 1 shows the advantages and disadvantages.

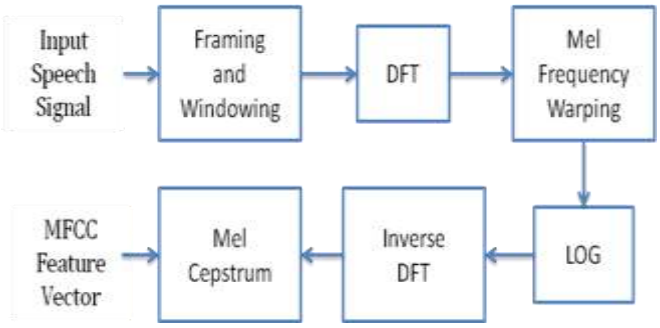

Fig. 2 Steps involved in MFCC Feature extraction

Table 1: Advantages and disadvantages of Feature Extraction techniques

Feature extraction technique	Advantages	Disadvantages
MFCC	• Provides good discrimination • Low correlation between coefficients • Not based on linear characteristics; hence, similar to the human auditory perception system • Important phonetic characteristics can be captured	• Low robustness to noise • In a continuous speech environment, a frame may not contain information of only one phoneme, but of two consecutive phonemes • Not flexible since the same basic wavelets have to be used for all speech signals
LPC	• Spectral envelope is represented with low dimension feature vectors • Good source-to-vocal tract separation is obtained • LPC method is simple to implement and mathematically precise	• Feature components are highly correlated • Cannot include a priori information on the speech signal under test

C. Speech recognition approaches

In the earlier years, dynamic programming techniques have been developed to solve the pattern-recognition problem [6].Subsequent researches were based on Artificial Neural Network (ANN) techniques, in which the parallel computing found in biological neural systems is mimicked. More recently, stochastic modeling schemes have been incorporated to solve the speech recognition problem, such as the Hidden Markov Modeling (HMM) approach. At present, much of the recent researches on speech recognition involve recognizing continuous speech from a large vocabulary using HMMs, ANNs, or a hybrid form [6]. These techniques are briefly explained below.

1) Template-Based Approaches

Template based approaches to speech recognition have provided a family of techniques that have advanced the field considerably during the last two decades. The underlying idea of this approach is simple. It is a process of matching unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match (Rabiner et al.,1979). This has the advantage of using perfectly accurate word models; but it also has the disadvantage that the pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. Template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred

words. This method was rather inefficient in terms of both required storage and processing power needed to perform the matching. Template matching was also heavily speaker dependent and continuous speech recognition was also impossible.

2) Knowledge-Based Approaches

The use of knowledge/rule based approach to speech recognition has been proposed by several researchers and applied to speech recognition (De Mori & Lam, 1986; Alikawa,1986; Bulot & Nocera, 1989), speech understanding systems (De Mori and Kuhn, 1992). The "expert" knowledge about variations in speech is hand-coded into a system. It uses set of features from the speech, and then the training system generates set of production rules automatically from the samples. These rules are derived from the parameters that provide most information about a classification. The recognition is performed at the frame level, using an inference engine (Hom, 1991) to execute the decision tree and classify the firing of the rules. This has the advantage of explicitly modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach was judged to be impractical, and automatic learning procedures were sought instead.

3) Neural Network-Based Approaches

Another approach in acoustic modeling is the use of neural networks. They are capable of solving much more complicated recognition tasks, but do not scale as excellent as Hidden Markov Model (HMM) when it comes to large vocabularies. Rather than being used in general-purpose speech recognition applications they can handle low quality, noisy data and speaker independence [7]. Such systems can achieve greater accuracy than HMM based systems, as long as there is training data and the vocabulary is limited. A more general approach using neural networks is phoneme recognition. This is an active field of research, but generally the results are better than HMMs. There are also NN-HMM hybrid systems that use the neural network part for phoneme recognition and the HMM part for language modeling.

4) Dynamic Time Warping (DTW)-Based Approaches

Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed [10]. A well known application has been ASR, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of HMM. In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. This technique is quite efficient or isolated word recognition and can be modified to recognize connected word also.

5) Hidden Markov Model (HMM)-Based Speech Recognition

The reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use [12] [13]. HMMs to represent complete words can be easily constructed (using the pronunciation dictionary) from phone HMMs and word sequence probabilities added and complete network searched for best path corresponding to the optimal word sequence. HMMs are simple networks that can generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state with, usually, mixtures of multivariate Gaussian distributions (the state output distributions). The parameters of the model are the state transition probabilities and the means, variances and mixture weights that characterize the state output distributions. Each word, or each phoneme, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM [16] for the separate words and phonemes. Current HMM-based large vocabulary speech recognition systems are often trained on hundreds of hours of acoustic data. The word sequence and a pronunciation dictionary and the HMM [7],[8] training process can automatically determine word and phone boundary information during training. This means that it is relatively straightforward to use large training corpora. It is the major advantage of HMM which will extremely reduce the time and complexity of recognition process for training large vocabulary.

Advantages and disadvantages of some most widely used modeling technique used in speech recognition are given in table 2.

Table 2: Advantages and disadvantages of classification techniques

Classification technique	Advantages	Disadvantages
HMM	• Simple to adapt • Able to model time distribution of speech signals • Inputs can be of variable length	• Based on the assumption that the probability of being in a particular state is dependent only on its preceding state • Emission probabilities are arbitrarily chosen
ANN	• Emission probabilities are arbitrarily chosen • Highly adequate for pattern recognition applications • Self-organizing • Self-learning • Self-adaptive in new environments	• Prone to over training a local minima problems

D. Performance evaluation of ASR techniques

The Performance of a speech recognition system is measurable. Perhaps the most widely used measurement is accuracy and speed. Accuracy is measured with the Word Error Rate (WER), whereas speed is measured with the real time factor. WER can be computed by the equation (2)

$$WER = S+D+I/N \quad (2)$$

Where S is the number of substitutions, D is the number of the deletions, I is the number of the insertions and N is the number of words in the reference.

The speed of a speech recognition system is commonly measured in terms of Real Time Factor (RTF). It takes time P to process an input of duration I. It is defined by the formula (3)

$$RTF = P/I \quad (3)$$

The comparison of the various speech recognition research based on the dataset, feature vectors, and speech recognition technique adopted for the particular language are given in the table 3.

Table 3: Comparison of various future extraction techniques with recognition techniques for speech data

Sr.No	Year	Author	Research Work	Data Used	Feature Extraction Technique	Recognition Technique	Language	Accuracy
1	2014	Preeti Kumari, D.Shakina Deiv, Mahua Bhattacharya	Native Hindi ASR and Bengali Accented Hindi ASR reconditioned	Large native Hindi and Bengali accented Hindi test speech.	Mel-frequency Cepstral coefficients (MFCC).	Hidden Markov Model (HMM)	Hindi	Native Hindi ASR - 94.93% Bengali Hindi ASR - 90.36 %

2	2015	Kasiprasad Mannepalli, Panyam Narahari Sastry, Maloji Suman	The recognition of coastal Andhra accent -Telugu speech	Small Vocabulary Speaker independent Continuous speech	MFCC	Gaussian mixture model (GMM)	Telugu	91%
3	2012	Ishan Bhardwaj, Student and Narendra D Londhe	The isolated word recognition for Hindi	speech is performed in speaker dependent, multi speaker and speaker independent manner with large vocabulary	MFCC	HMM	Hindi	Speaker dependent-99%, multi speaker-98% and speaker independent -97.5
4	2014	Pialy Barua, Ainul Anam Shahjamal Khan, Muhammad Sanaullah	Reorganization of speech from Bangla commands.	Two native Bangla commands used	MFCC	ANN	Bangla	83%
5	2015	S. Ananthi and P. Dhanalakshmi	Convert spoken word into text	Large vocabulary	LPC and MFCC	HMM and Support Vector Machine (SVM)	English	SVM-MFCC accuracy of 91.46% and HMM - MFCC accuracy of 98.92%
6	2016	Hiral B. Chauhan and B.A. Tanawala	Performance Based Comparison for Gujarati Numbers Detection	small Vocabulary	MFCC and LPC	VQ algorithm	Gujarati - Numbers	LPC - 86 % and MFCC 98 %
7	2016	Kartiki Gupta, Divya Gupta	Comparison of LPC MFCC and RASTA	Large vocabulary	LPC, MFCC and RASTA	Matches the words of the input signal to words in database.	English	LPC -91.4 % RASTA - 94.27 % MFCC -99.9 %

III. CONCLUSION AND FUTURE SCOPE

In this review, we have discussed the different feature extraction and classifier techniques of speaker identification through previous experimental research and state its advantages and disadvantages. We also presented the comparative analysis table for speech recognition techniques used and found that MFCC is used widely for feature extraction of speech and HMM and ANN are best among all modeling technique. Finally it is also observed that very less work has been done on Indian languages specially local languages like Gujarati.

The future scope of this study is to develop the complete accurate applications and will focus on recognition of Gujarati speech with the MFCC and HMM/ANN as a hybrid approach for classification for better accuracy.

ACKNOWLEDGMENT

We thank our colleagues from L. D. College of engineering, Ahmedabad who provided insight and expertise that greatly assisted the survey

REFERENCES

[1] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, "Comparative study of automatic speech recognition techniques" IEEE-IET Signal Process., 2013, Vol. 7, Iss. 1, pp. 25–46
[2] Vimala, C., Radha, V.: 'A review on speech recognition challenges and approaches', World Comput. Sci. Inf. Technol., 2012, 2, (1), pp. 1–7
[3] Goutam Saha, Ulla S. Yadhunandan " Modifield Mel-Frequency Cepstral coefficient Department of Electronics and Electrical communication Engineering India Institute of Technology ,Kharagpur Kharagpur-721302 West Bengal,India.
[4] DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03$17.00 ©2003 IEEE
[5] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory andEngineering, Vol.2, No.6, December, 2010 1793-8201.
[6] Zhao Lishuang , Han Zhiyan, "Speech Recognition System Based on Integrating feature and HMM", 2010 International Conference on Measuring Technology and Mechatronics Automation, 978-0-7695-3962-1/10 $26.00 © 2010 IEEE
[7] Vimal Krishnan V. R, Athulya Jayakumar and Babu Anto.P, "Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network", 4th IEEE International Symposium on Electronic Design, Test & Applications, 0-7695-3110-5/08 $25.00© 2008 IEEE
[8] Preeti Kumari I, D.Shakina Deiv, Mahua Bhattacharya, "Automatic Speech Recognition of Accented Hindi Data", 2014 INTERNATIONAL CONFERENCE ON COMPUTATION OF POWER, ENERGY, INFORMATION AND COMMUNICATION (ICCPEIC) –IEEE
[9] Kasiprasad Mannepalli,Panyam Narahari Sastry, Maloji Suman "MFCC-GMM based accent recognition system for Telugu speech signals", Springer Science+Business Media New York 2015, DOI 10.1007/s10772-015-9328-y
[10] Ishan Bhardwaj, Narendra D Londhe ," Hidden Markov Model Based Isolated Hindi Word Recognition", 2012-IEEE 2nd International Conference on Power, Control and Embedded Systems
[11] Pialy Barua, Ainul Anam Shahjamal Khan, Muhammad Sanaullah, "Neural Network Based Recognition of Speech Using MFCC Features" 3rd INTERNATIONAL CONFERENCE ON INFORMATICS, ELECTRONICS & VISION 2014 @IEEE
[12] Pialy Barua, Ainul Anam Shahjamal Khan, Muhammad Sanaullah, "Neural Network Based Recognition of Speech Using MFCC Features" 3rd INTERNATIONAL CONFERENCE ON INFORMATICS, ELECTRONICS & VISION 2014 @IEEE
[13] Kartiki Gupta, Divya Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech Recognition System", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence) 978-1-4673-8203-8/16/$31.00 c 2016 IEEE pp 493
[14] S. Ananthi and P. Dhanalakshmi," SVM and HMM Modeling Techniques for Speech Recognition Using LPCC and MFCC Features", Springer International Publishing Switzerland, DOI: 10.1007/978-3-319-11933-5_58
[15] J. Ashraf, N. Iqbal, N. S. Khattak, and A. M. Zaidi, "Speaker independent urdu speech recognition using hmm," in Informatics and Systems (INFOS), 2010 The 7th International Conference on, pp. 1–5, IEEE, 2010.
[16] Y.-H. B. Chiu and R. M. Stern, "Minimum variance modulation filter for robust speech recognition," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pp. 3917–3920 , IEEE, 2009.
[17] G. Muhammad, Y. Alotaibi, M. N. Huda, et al., "Automatic speech recognition for bangla digits," in Computers and Information Technology, 2009. ICCIT'09. 12th International Conference on, pp. 379–383 , IEEE, 2009.
[18] Mousmita Sarma and Kandarpa Kumar Sarma, "Speech Recognition in Indian Languages—A Survey", © Springer India 2015 K.K. Sarma et al. (eds.), Recent Trends in Intelligent and Emerging Systems, Signals and Communication Technology, DOI 10.1007/978-81-322-2407-5_14.