KIET KIET International Journal of Communications &Electronics
Volume. No. 2, Issue No. 2, July – Oct 2014, ISSN: 2320 - 8996

A Review on Speech Feature Extraction and Speech Feature Matching Technique

Mr. Mahesh Kumar Patil[1], Prof. Dr. (Mrs.) L.S. Admuthe[2], Mr. Prashant P Zirmite[3,] Prof. Mr. N.B. Kapase[4]
Assistant Professor, Electronics Engineering, Textile & Engineering Institute, Ichalkaranji

Abstract- **Speech is one of the natural forms of communication. Different speech sounds can be characterized by set of spectral and temporal properties which depend on the speech features such as speech waveform or speech spectrum. The speech recognition system contains two main tasks- Feature Extraction and Feature Matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each spoken word. Feature matching involves the actual procedure to identify the actual spoken words by comparing extracted features from set of known database.**

Keywords- **MFCC, LPC, HMM, DTW, Modeling, Testing.**

I. Introduction

Humans are fairly good at identifying speakers based on their voices alone. The large amount of work in the field of speaker recognition over the previous 30 years has been predicated on the belief that automated systems ought to be able to do as well, or even better, than humans. Yet we still lack a solid understanding of those characteristics of speech that index an utterance as originating in one speaker rather than another.

The general area of speaker recognition encompasses two fundamental tasks: speaker identification and speaker verification. Speaker identification is the task of assigning an unknown voice to one of the speakers known by the system: it is assumed that the voice must come from a fixed set of speakers. Thus, the system must solve a n-class classification problem and the task is often referred to as closed-set identification. On the other hand, speaker verification refers to the case of open-set identification: it is generally assumed that the unknown voice may come from an impostor. Regardless of the specific task at hand, it is common practice to adopt a probabilistic approach that predicts the likelihood that a given speech sample belongs to a given speaker. The base system for speaker recognition is usually composed of a speech parameterization module and a statistical modeling module which are responsible for the production of a machine readable parameterization of the speech samples and the computation of a statistical model from the parameters. The main difference between speaker identification and speaker verification is that in the first case the system provides one model for each speaker, while, in the second case, the system provides a total of two models: one for the hypothesized speaker and one representing the hypothesis that the speech sample comes from some other speaker—the background model.

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. The speaker recognition system may be viewed as working in a four stages.

A. Analysis
B. Feature extraction
C. Modeling
D. Testing/Matching techniques

A. Speech analysis

In speech analysis technique Speech data contains different types of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The physical structure and dimension of vocal tract as well as excitation source are unique for each speaker. The speech analysis deals with stages with suitable frame size for segmenting speech signal for further analysis and extracting [2]. The speech analysis is 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. Studies have been made in using 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 mainly to analyze and extract the characteristic of the excitation state. The excitation source information is relatively fast varying compared to vocal tract information, so small frame size and shift are required to best capture the speaker-specific information [3].

3) Supra-segmental Analysis: In this case, speech is analyzed by using the frame size and shift of 100-300 ms to extract speaker information mainly due to behavioral tract and speech is analyzed using the frame size.

This technique is used mainly to analyze and characteristic due to behavior character of the Speaker. These include word duration, intonation, speaker rate, accent etc.

B. Feature extraction techniques

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. The utterance 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:

i. Easy to measure extracted speech features
ii. It should not be susceptible to mimicry
iii. It should show little fluctuation from one speaking environment to another
iv. It should be stable over time
v. 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. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. The figure1 shows the steps involved in LPC feature extraction.

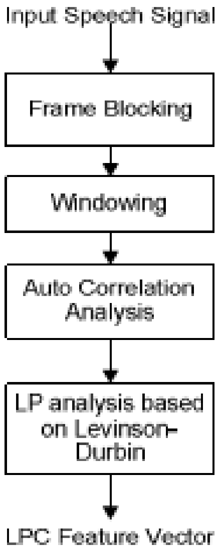
Input Speech Signal
Frame Blocking
Windowing
Auto Correlation Analysis
LP analysis based on Levinson-Durbin
LPC Feature Vector

Fig 1. Steps involved in LPC Feature Extraction

2) Mel Frequency Cepstral Coefficients (MFCC) :

The following figure 2 shows the steps involved in MFCC feature extraction. 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 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.

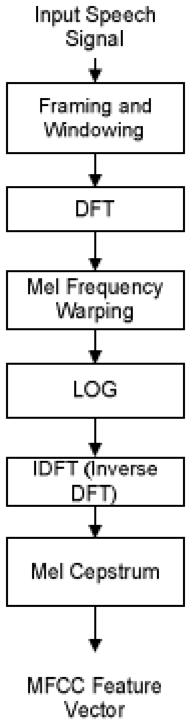

Fig. 2 Steps involved in MFCC Feature Extraction

It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula (1).

$$Mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \qquad (1)$$

C. Modeling technique

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided into two classification speaker recognition and speaker identification. The speaker identification technique automatically identifies who is speaking on basis of individual information integrated in speech signal. The system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be dividing into two methods, text- dependent and text independent methods. In text dependent method the speaker say key words or sentences having the same text for both training and recognition trials. Whereas text independent does not rely on a specific texts being spoken [7]. Following are the modeling which can be used in speech recognition process:

1) The acoustic-phonetic approach

This method is indeed viable and has been studied in great depth for more than 40 years. This approach is based upon theory of acoustic phonetics and postulates. The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach (Hemdal and Hughes 1967). Which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time? Even though, the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds (the so-called co articulation effect), it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine [10].There are three techniques that have been applied to the language identification. Problem phone recognition, Gaussian mixture modeling, and support vector machine classification. The acoustic phonetic approach has not been widely used in most commercial applications.

2) Pattern Recognition approach

The pattern-matching approach (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns [6].

2) *Template based approaches*

Template based approaches matching (Rabiner et al., 1979) unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match. This has the advantage of using perfectly accurate word models. Recognition is carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. Usually templates for entire words are constructed. One key idea in template method is to derive a typical sequence of speech frames for a pattern (a word) via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another key idea is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker. [7].

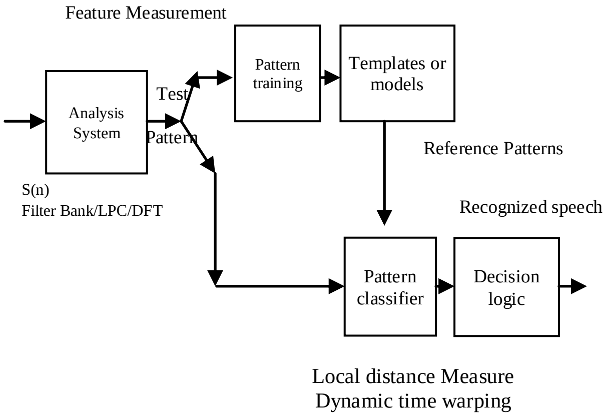

Fig.3 Block diagram of Pattern recognition

3) *Dynamic Time Warping (DTW)*

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she was walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed. Any data which can be turned into a linear representation can be analyzed with DTW. 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. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models. Continuity is less important in DTW than in other pattern matching algorithms. This technique is quite efficient for isolated word recognition and can be modified to recognize connected word also [8].

5) Knowledge Based Approach Knowledge

Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. Vector Quantization (VQ) [9] is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. The utility of VQ here lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure. Knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms. Algorithms enable us to solve problems. Knowledge enables the algorithms to work better. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

6) The Artificial Intelligence Approach

The Artificial Intelligence approach [10] is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features.

A large body of linguistic and phonetic literature provided insights and understanding to human speech processing this knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques, such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms.

7) Statistical Based Approach

In this approach, variations in speech are modeled statistically (e.g., HMM), using automatic learning procedures. This approach represents the current state of the art. Modern general-purpose speech recognition systems are based on statistical acoustic and language models. Effective acoustic and language models for ASR in unrestricted domain require large amount of acoustic and linguistic data for parameter estimation. Processing of large amounts of training data is a key element in the development of an effective ASR technology now a days. The main disadvantage of statistical models is that they must make *a priori* modeling assumptions, which are liable to be inaccurate, handicapping the system's performance.

This new approach is a radical departure from the current HMM-based statistical modeling approaches. For text independents speaker recognition use left-right HMM for identifying the speaker from simple data and also HMM having advantages based on Neural Network and Vector Quantization. The HMM is popular statistical tool for modeling a wide range of time series data. In Speech recognition area, HMM have been applied with great success to problem as part of speech classification [11]. The K-means algorithm is also used for statistical and clustering algorithm of speech Based on the attribute of data .The K in K-means represents the number of clusters the algorithm should return in the end. As the algorithm starts K points known as cancroids are added to the data space. The K-means algorithm is a way to cluster the training vectors to get feature vectors. In this algorithm clustered the vectors based on attributes into k partitions. It uses the k means of data generated from Gaussian distributions to cluster the vectors. The objective of the k-means is to minimize total intra-cluster variance.

8) Stochastic Approach

Stochastic modeling [12] entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability's, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state Markov model and a set of output distributions. The transition parameters in the Markov chain models, temporal variability's, while the parameters in the output distribution model, spectral variability's. These two types of variability's are the essence of speech recognition. Compared to template based approach,

Hidden Markov modeling is more general and has a firmer mathematical foundation. A template based model is simply a continuous [12].

D. MATCHING TECHNIQUES

Speech-recognition engines match a detected word to a known word using one of the following techniques (Svendsen et al., 1989)

1) Whole-word matching:

The engine compares the incoming digital-audio signal against a pre-recorded template of the word This technique takes much less processing than sub-word matching, but it requires that the user (or someone) prerecord every word that will be recognized - sometimes several thousand words. Whole-word templates also require large amounts of storage (between 50 and 512 bytes per word) and are practical only if the recognition vocabulary is known when the application is developed [13].

2) Sub-word matching:

The engine looks for sub-words – usually phonemes and then performs further pattern recognition on those. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes perword). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand [14] [15].

TABLE I
RESULTS OBTAINED USING DIFFERENT FEATURE EXTRACTION AND MATCHING TECHNIQUE

Author	Year	Research Work	Nature of Data	Feature Extraction Technique	Feature Matching Technique	Language	Accuracy
Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda	2009	Automatic Speech Recognition for Bangia Digits	Small vocabulary Speaker independent Isolated digit	Mel-Frequency Cepstral Coefficients (MFCCs)	Hidden Markov Model (HMM)	Bangia	more than 95% for digits (0-5) and less than 90% for digits (6-9)
Corneliu Octavian Dumitru, Inge Gavat	2006	A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language	Large vocabulary Speaker independent Continuous speech	PLP, MFCC, LPC	Hidden Markov Models (HMM)	Romanian	MFCC-90,41%, LPC-63,55%. and PLP 75,78%
Bassam A. Q. Al-Qatab , Raja N. Ainon	Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)	MFCC	HMM	Arabic	97.99%	Bassam A. Q. Al-Qatab , Raja N. Ainon	Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)
M.Chandrasekar, and M.Ponnavaiko	2008	Tamil speech recognition: a complete model	Medium vocabulary Speaker dependent Isolated Speech	MFCC	Back-Propagation Network	Tamil	80.95 %

III. PERFORMANCE OF A SYSTEM

For the performance analysis of the system, the following parameters will be considered for accuracy and speed measurement of the system.

1. Word Error Rate : $WER = (S + D + I)/N$
 Where S is the number of substitutions, D is the number of deletions and I is number of insertions and N is the number of words in the reference speech sample.
2. Word Recognition Rate : *WRR= 1- WER*
3. Real Time Factor : $RTF = P/I$ It takes time P to process an input of duration *I.*

IV. CONCLUSION

In this review, the fundamentals of speech recognition are discussed and its recent progress is investigated. The various approaches available for developing an ASR system are clearly explained with its merits and demerits. The performance of the ASR system based on the adopted feature extraction technique and the speech recognition approach for the particular language is compared in this paper. In recent years, the need for speech recognition research based on large vocabulary speaker independent continuous speech has highly increased. Based on the review, the potent advantage of HMM approach along with MFCC features is more suitable for these requirements and offers good recognition result.

FUTURE SCOPE

These techniques will enable us to create increasingly powerful systems, deployable on a worldwide basis in future. Results indicate that there is still room for recognition rate improvements. Using ANN technique these results can be increased further. Future work will focus on better selection of word groups and using speaker-dependent word groups.

REFERENCES

[1] Lawrence Rabiner, Biing Hwang Juang, Fundamental of Speech Recognition, Copyright 1999 by AT&T.

[2] GIN-DER WU AND YING LEI " A Register Array based Low power FFT Processor for speech recognition" Department of Electrical engineering national Chi Nan university Puli545 Taiwan.

[3] B. Yegnanarayana, S.R.M. Prasanna, J. M. Zachariah, and C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed- text specker verification system," IEEE Trans. Speech Audio Process.,vol.13(4), pp. 575-82, July2005.

[4] 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 and Engineering, Vol.2, No.6, December, 2010 1793-8201.

[5] Satyanarayana "short segment analysis of speech for enhancement" institute of IIT Madras feb.2009

[6] C.S.Myers and L.R.Rabiner, A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition, IEEE Trans. Acoustics, Speech Signal Proc.,ASSP-29:284-297,April 1981.

[7]H.Sakoe and S.Chiba, Dynamic programming algorithm optimization for spoken word recognition ,IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26(1).1978

[8] Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010

[9] Keh-Yih Su et.al, Speech Recognition using weighted HMM and subspace IEEE Transactions on Audio, Speech and Language.

[10] R.K.Moore, Twenty things we still don t know about speech, Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology, 1994.

[11] Shigeru Katagiri et.al, A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization, IEEE Transactions on Audio Speech and Language processing Vol.1, No.4

[12] M.Weintraub et.al, linguistic constraints in hidden markov Model based speech recognition, Proc.ICASSP, pp.699-702, 1989.

[13]S.katagiri, Speech Pattern recognition using Neural Networks.

[14]L. R .Rabiner and B.H.jaung," Fundamentals of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersy, 1993

[15] D. R. Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave, Tech. Report No.C549, Computer Science Dept., Stanford Univ., September 1966.