Speech Feature Extraction and Classification: A Comparative Review

Akansha Madan
M.tech Student
Amity University, Noida

Divya Gupta
Assistant Professor
Amity University, Noida

ABSTRACT

This paper gives a brief survey on speech recognition and presents an overview for various techniques used at various stages of speech recognition systems. Researchers has been working in this research area for many years however accuracy for speech recognition still attention for variation of context, speaker's variability, environment conditions .The development of speech recognition system requires certain concepts to be included-Defining different classes of speech, techniques for speech feature extraction, speech classification modeling and measuring system performance .The main aim of this paper is to discuss and compare different approaches used for feature extraction and classification stages in speech recognition system.

Keywords

Speech Recognition, Robust speech recognition, Speech feature extraction, Classification

1. INTRODUCTION

Speech is the principal source of communication among humans to show their ideas, feelings and thoughts to each other. In fact, using speech as a source for controlling one's surroundings is always an intriguing concept. Speech recognition technology has made it possible for computer to listen human voice commands and interpret human languages. Speech recognition is the process of converting a given input signal into sequence of words, by means an algorithm implemented as a computer program.

In other words, Speech Recognition system allows a computer to identify the words that a person speaks into a microphone or telephone and convert it into readable text. The speech recognition system would support many valuable applications that require human interaction with machine [13].

1.1 Classification of speech recognition systems based on speech

Speech recognition systems can be divided into different classes based on type of speech utterances they are capable to recognize. Different classes are described as below [20]:

a) Isolated Words- These systems require each utterance to contain lack of an audio signal on both sides of constructed sample window. It accepts single word at a given time.

b) Connected Words- These systems allows separate utterances to be spoken together with a minimum pause between them.

c) Continuous Speech- These systems allows people to speak almost in a natural way. Continuous speech recognizers are difficult to create as they require more effort to determine word boundaries.

d) Spontaneous Speech- This type of speech is natural and not rehearsed. Speech recognition system with spontaneous speech should be able to handle different natural speech features such as words being run together including slight stutters. Spontaneous speech may include mispronunciations, false-starts, and non-words.

1.2 Overview of speech recognition system

Development in speech technology has been inspired by the reason that people desire to develop mechanical models that permits the emulation of human verbal communication capabilities. Speech processing allow computer to follow voice commands and different human languages.

Fig 1 depicts a basic model of speech recognition system that represents different stages of a system including pre-processing, speech feature extraction, classification and language model [22].

The pre-processing transforms the input signal before any information can be extracted at feature extraction stage. After pre-processing, feature extraction stage extracts necessary vectors to be used at modeling stage. These vectors must be robust to noise for better accuracy.

Classification stage recognize the speech text using extracted features and language model where Language Model contains syntax and semantics related to language responsible that helps classifier to recognize the input utterance[22].

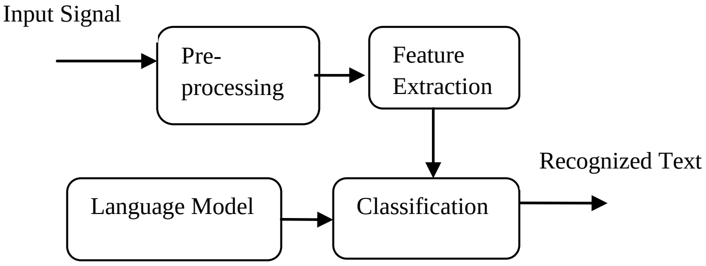

Fig1. Stages of Speech Recognition

2. FEATURE EXTRACTION FOR SPEECH RECOGNITION

Speech feature extraction is responsible for transformation of the speech signals into stream of feature vectors coefficients which contains only that information which is required for the identification of a given utterance. As every speech has different unique attributes contained in spoken words these attributes can be extracted from a wide range of feature extraction techniques and can be employed for speech recognition task. But extracted feature should meet certain criteria while dealing with the speech signal such as: extracted speech features should be measured easily, extracted features should be consistent with time, and features should be robust to noise and environment [18].

The feature vector of speech signals are typically extracted using spectral analysis techniques such as Mel- frequency cepstral coefficients, linear predictive coding wavelet transforms. The most widely used feature extraction techniques along with their merits and demerits are discussed below:

a. Linear Predictive Coding

One of the often applied techniques at feature extraction stage for signal analysis is the method of linear prediction that derives the fundamental parameters of speech. It is a time domain methodology that analyzes the human tract structure to provide a precise estimate of the speech parameters when words are spoken.

The main goal of this method is the approximation of a given speech sample as a linear combination of earlier speech samples. By reducing the sum of squared differences (over a finite interval)to its minimum value between the actual speech samples and calculated values, a unique parameter set or predictor coefficients can be found. The coefficient values predicted forms the building block for LPC of speech [14].The following figure 2 shows the steps involved in LPC feature extraction [18]

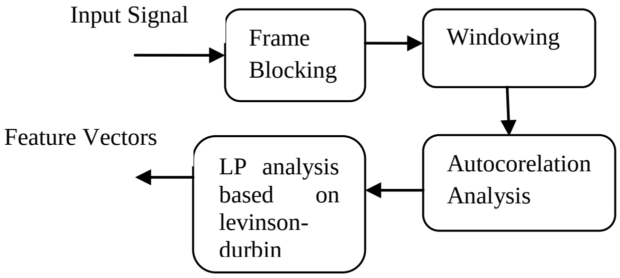

Fig 2. Stages of Linear Predictive Coding

b. Mel Frequency Cepstral Coefficients (MFCC)

Many developers in speech recognition apply MFCC at feature extraction stage. The MFCC tries to mimic the human ear, where frequencies are nonlinearly resolved across the audio spectrum. Hence, the purpose of the Mel filters is to distort the frequency such that it obey the spatial relationship of the hair cell distribution of the human ear. Hence, the mel frequency scale corresponds to a linear scale below 1 kHz, and algorithmic scale above the 1 kHz. [17].

MFCC methodology is based on the short-term analysis, where feature vector is computed from each frame separately. The coefficients are extracted by taking speech sample as an input and then hamming window is applied to reduce the discontinuities of a signal. Then the Mel filter bank is generated by applying FFT. According to Mel frequency warping, as width of the triangular filters differs therefore log total energy in a critical band around the center frequency is combined. The numbers of coefficients are then obtained after wrapping. In last, the Inverse Discrete Fourier Transformer is applied for the calculation of cepstral coefficients [3] [1]. It transforms the log of the domain coefficients to the frequency domain where N is the length of the FFT. MFCC can be computed by using the formula [8, 18].

Mel (g) = 2595*log10 (1+g/700)
The following figure3 shows the steps involved in MFCC[18]

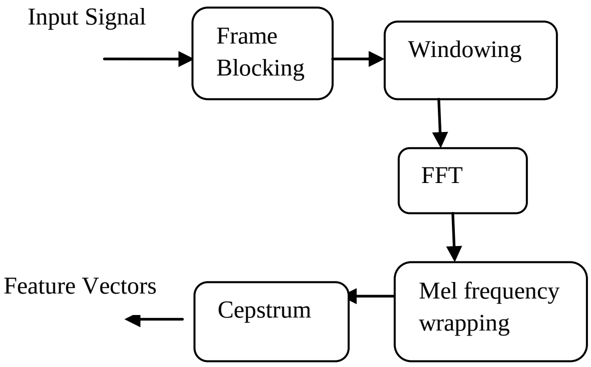

Fig 3. Stages of MFCC

Table 1: Comparison of techniques is given below:

Feature Extraction Technique	Merits	Demerits
MFCC	a. gives good distinction[15] b. little association among coefficients [15] c. not based on linear attributes; hence, identical to the human auditory perception system [15, 11] d. necessary phonetic attributes can be gathered [15]	a. little noise robustness [15, 10] b. in a continuous speech environment, a frame may not contain information of t of two consecutive phonemes rather than only one phoneme [4, 10] c. As it takes into account only the power spectrum and ignores the phase spectrum, hence it provides limited representation.
Linear Predictive Coding	a. Low dimension feature vectors is used to present spectral envelope [15,16] b. good source-to-vocal tract separation is possible[15] c. LPC method can be implemented easily and provides precise mathematical representation.[15]	a. linear scales are not enough for speech production or perception representation [14] b. Highly associated Feature components [15] c. unable to include a priori information for speech signal under test [22]

3. CLASSIFICATION FOR SPEECH RECOGNITION

Several developers are working to develop suitable classifier that correctly identifies speech utterances over different conditions. In speech recognition, a supervised pattern classification system has been trained using examples with appropriate labels. The alternate way for training pattern classifiers is by unsupervised learning.

After, a proper representation has been found through feature extraction approaches, a classifier can be developed by using any of available approaches [21]. The decision regarding classifier to be used has been a tough task and should be chosen considering the issues e.g. availability of classifier, knowledge of classifier, amount of data required to train it [13]. Different classification models are summarized in the table below [22]

Table 2: comparison of classification approaches

Classification Technique	Merits	Demerits
Hidden Markov Models	a. Models time distribution of speech signals [5]. b. Simple to develop [22]. c. Can model discrete or continuous symbols [7]. d. Support variable length input [6].	a. Assumes that probability to exist in a particular state is dependent on its previous state [22].
Artificial Neural Network	a. Self-organizing and self- learning ability [12, 16]. b. Easily adjustable to new environments and robust [12, 16]. c. Suitable for pattern recognition [12].	a.Requires over training of data [5, 12].
Support Vector Machine	a. It does not face problems like local minima and over-training. b. Able to deal with high – dimensional input vectors and robust[5].	a. Does not support variable length input. They need to be fixed length [5]. b. As number of classes increases, computational cost also increases [5]. c. Not capable to deal with large databases [5].

4. OUTLINE OF FEATURE EXTRACTION AND CLASSIFICATION STAGES

Table 3: Outline of feature extraction and classification stages

Concept	**Feature Extraction**	**Classification**
Definition	It is process where speech signal is converted into sequence of feature vectors coefficients which contains only necessary information required for speech recognition [13].	It is the process of mapping feature vectors found using language model to recognize the input utterance [23].
Working Principle	The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectro temporal analysis of the signal and extracts natural attributes depicting the envelope of the power spectrum for short speech intervals. The second stage gathers an extended feature vector composed of static as well as dynamic features. Finally, the last stage (if exist) converts these enhanced feature vectors into small-scale and robust vectors that are then provided to the recognizer [21].	The classification stage can work using two different approaches. The first method is known as generative approach, where the joint probability distribution has been identified with given observations and the class labels .The second approach is known as discriminative approach which finds the conditional distribution using a parametric model, where the parameters are identified using training set comprising of pairs of the input vectors and their corresponding target output vectors.[23]
Property	Feature vectors must have ability to differentiate among classes and should be robust to environment conditions such as noise [22].	Classification stage classify input speech signal according to extracted features. It allows feature detection and labeling classes for recognition of input signal.
Techniques used	MFCC ,linear predictive coding, principal component analysis, linear discriminant analysis, wavelet transform	Hidden Markov Model, ,SVM, ANN
comment	Work well with clean environment and isolated digits.Fusion of different techniques and noise elimination techniques can be used for higher accuracy in recognition.	When training and testing acoustic conditions differ, the accuracy of the systems rapidly degrades. And also computational cost increases as training data increases. Different models can be fused together to take into account time variation of signals with reduction data required for training.

5. PERFORMANCE MEASUREMENT OF SPEECH RECOGNITION APPROACHES

The performance of a speech recognition system is measurable in terms of their accuracy and speed. Accuracy can be measured in terms of Word Error Rate (WER), Command Success Rate (CSR).

WER is the commonly used metric for measuring performance of speech recognition systems. Word Error Rate is computed by the equation [3] as given below

$$WER = \frac{S+D+I}{N}$$

Where S is no. of substitutions, D is number of deletions, I is number of insertions and N is number of words in 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 [1] as given below

$$RTF = \frac{P}{I}$$

6. CONCLUSION AND FUTURE DIRECTION

In this paper, the fundamental concepts of speech recognition are discussed along with widely used feature extraction and classification techniques used for speech recognition. The various approaches available for developing speech recognition system are compared along with their merits and demerits.

On the basis of review, as feature extraction is the important stage of speech recognition and MFCC work well only for clean environment. Thus, other feature extraction techniques can be combined with MFCC to be used for robust speech recognition. And to allow MFCC to work well noise can be eliminated using Empirical Mode Decomposition in future.

The performance of the system with above mention approaches can then be compared on different scenarios like noise corrupted connected digits as future work. Discrete hidden markov model can be used for classification for better word recognition as they consider time distribution of speech signals.

7. REFERENCES

[1] 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.

[2] O'Shaughnessy, D,"Interacting with computers by voice: automatic speech recognition and synthesis', Proc. IEEE, 2003, 91, (9), pp. 1272–1305.

[3] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.

[4] Muller, D.N., de Siqueira, M.L., Navaux, P.O.A.: 'A connectionist approach to speech understanding'. Int. Joint Conf. on Neural Networks, 2006 (IJCNN'06), Vancouver, BC, 2006, pp. 3790–3797

[5] Solera-Urena, R., Padrell-Sendra, J., Martin-Iglesias, D., Gallardo-Antolin, A., Pelaez-Moreno, C., Diaz-De-Maria, F.: 'SVMs for automatic speech recognition: a survey', Progress in nonlinear speech processing (Springer-Verlag, Berlin, Heidelberg, 2007), pp. 190–216.

[6] Milone, D.H., Di Persia, and L.E.: 'Learning hidden Markov models with hidden Markov trees as observation distributions'. Ninth Argentine Symp. Artificial Intelligence (ASAI 2007), Mar del Plata, Argentina, 2007, pp. 13–22.

[7] Mporas, I., Ganchev, T., Siafarikas, M., Fakotakis, Comparison of speech features on the speech recognition task', J. Computer. Sci., 2007, 3, (8), pp. 608–616

[8] M. Chandrasekar, M. Ponnavaikko, "Tamil speech recognition: a complete model", Electronic Journal «Technical Acoustics» 2008, 20.

[9] Korba, M.C.A., Messadeg, D., Djemili, R.H.B.: 'Robust speech recognition using perceptual wavelet denoising and Mel-frequency product spectrum cepstral coefficient features', Informatics, 2008, 32, pp. 283–288

[10] Xuefei, L.: 'A new wavelet threshold denoising algorithm in speech recognition '.Asia-Pacific Conf. on Information Processing, 2009(APCIP 2009), Shenzhen, 2009, pp. 310–313.

[11] Vimal Krishnan, V.R., Babu Anto, P.: 'Features of wavelet packet decomposition and discrete wavelet transform for Malayalam speechrecognition', Recent Trends Eng., 2009, 1, (2), pp. 93–96

[12] Zhou, P., Tang, L.Z., Xu, D.F.: 'Speech recognition algorithm of parallel sub band HMM based on wavelet analysis and neural network', Inf. Technol. J., 2009, 8, pp. 796–800.

[13] M.A.Anusuya, S.K.Katti, Speech Recognition by Machine: A Review, (IJCSIS) International Journal of Computer Science and Information Security, 2009, Vol. 6, No. 3.

[14] 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

[15] Anusuya, M., Katti, S.: 'Front end analysis of speech recognition: a review', Int. J. Speech Technol., 2011, 14, (2), pp. 99–145.

[16] Venkateswarlu, R.L.K., Kumari, R.V., Jayasri, G.V.: 'Speech recognition using radial basis function neural network'. Third Int. Conf. on Electronics Computer Technology (ICECT), 2011, Kanyakumari, 2011, pp. 441–445

[17] Sivaram, G.S.V.S., Hermansky, H.: 'Multilayer perceptron with sparse hidden outputs for phoneme recognition'. 2011 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Prague, 2011, pp. 5336–5339

[18] Vimala, C., Radha, V.: 'A review on speech recognition challenges and approaches', World Computer. Sci. Inf. Technol., 2012, 2, (1), pp. 1–7

[19] Ashok Shigli, Ibrahim Patel, A Spectral Feature Process for Speech Recognition Using HMM with MFCC Approach, National Conference on Computing and Communication Systems (NCCCS), 2012, 978-1-4673-1953-9

[20] Sharada C. Sajjan, Vijaya C, Comparison of DTW and HMM for Isolated Word Recognition, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 2012, 978-1-4673-1039, IEEE.

[21] Shing-Tai pan, Tzung-Pei Hong, Robust Speech Recognition by DHMM with A Codebook Trained by Genetic Algorithm, Journal of Information Hiding and Multimedia Signal Processing, October 2012, vol-3.

[22] Anjivani S. Bhabad, Gajanan K. Kharate, An Overview of Technical Progress in Speech Recognition, International Journal of Advanced Research in Computer Science and Software Engineering, March 2013,.Volume 3, Issue 3.

[23] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, Comparative study of automatic speech recognition techniques, IET Signal Process., 2013, Vol. 7, Iss. 1, pp. 2.