Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Speech Recognition: An Approach to Modernization

Deepali H. Shah
Associate Professor
Instrumentation & Control Department
L. D. College of Engineering, Ahmedabad,
Gujarat, India

Tejas V. Shah
Associate Professor
Instrumentation & Control Department
S. S. Engineering College, Bhavnagar,
Gujarat, India

***Abstract*— Speech is the major way of communication among human beings. Through speech man can communicate naturally and effectively. Speech recognition is a field of computer science that deals with designing computer systems that recognize spoken words. This paper presents the basic idea of speech recognition, types of speech recognition, different application areas and different approaches for recognizing the speech of the different speaker.**

***Keywords*— LPC, MFCC, RASTA, linguistic, phonetic**

I. INTRODUCTION

Speech Recognition is the process of converting a speech signal to a sequence of words, by means of algorithms implemented as a computer program. Speech is the most natural form of human communication. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages [1]. The primary function of the speech recognition engine is to process spoken input and translate it into text that an application understands. The application can then do one of two things [2]:

- The application can interpret the result of the recognition as a command. In this case, the application is a command and control application. An example of a command and control application is one in which the caller says "check balance", and the application returns the current balance of the caller's account.
- If an application handles the recognized text simply as text, then it is considered a dictation application. In a dictation application, if you said "check balance," the application would not interpret the result, but simply return the text "check balance".

For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years. Based on major advances in statistical modelling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, Data entry, voice dictation, access to information: travel, banking, Commands, Avionics, Automobile portal, speech transcription, Handicapped people (blind people) supermarket, railway reservations etc. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operator services [3].

II. CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

Speech recognition systems can be separated in several different classes by describing the type of speech utterance, type of speaker model, type of channel and the type of vocabulary that they have the ability to recognize [4].

A. Types of Speech Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences. The types of speech utterance are [4]:

1) *Isolated Words:* Isolated word recognizers usually require each utterance to have quiet (absence of voice)on both sides of the sample window. It accepts single words or single utterances at a time .This is having "Listen and Non Listen state".[3]This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs. It is comparatively simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The disadvantage of this type is choosing different boundaries affects the results [4].
2) *Connected Words:* Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them [4].

3) *Continuous Speech:* Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation. It includes a great deal of "co articulation", where adjacent words run together without pauses or any other apparent division between words. Continuous speech recognition systems are most difficult to create because they must utilize special methods to determine utterance boundaries. As vocabulary grows larger, confusability between different word sequences grows [4].

4) *Spontaneous Speech:* This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features such as words being run together and even slight stutters. Spontaneous (unrehearsed) speech may include mispronunciations, false-starts, and non-words [4].

B. Types of Speaker Model

Each speaker has special voice, due to his unique physical body and personality. Speech recognition system is classified into three main categories as follows

1) *Speaker dependent models:* Speaker dependent systems are developed for a particular type of speaker. They are generally more accurate for the particular speaker, but could be less accurate for other type of speakers. These systems are usually cheaper, easier to develop and more accurate. But these systems are not flexible as speaker independent systems [4], [5].

2) *Speaker Independent Models:* Speaker Independent system can recognize a variety of speakers without any prior training.. A speaker independent system is developed to operate for any particular type of speaker. It is used in Interactive Voice Response System (IVRS) that must accept input from a large number of different users. But drawback is that it limits the number of words in a vocabulary. Implementation of Speaker Independent system is the most difficult. Also it is expensive and its accuracy is lower than speaker dependent systems [4], [5].

3) *Speaker Adaptive Models:* Speaker adaptive speech recognition system uses the speaker dependent data and adapt to the best suited speaker to recognize the speech and decreases error rate by adaption. They adapt operation according to characteristics of speakers [5].

4) *Types of Vocabulary:* The size of vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. Some applications only require a few words (e.g. numbers only), others require very large dictionaries (e.g. dictation machines). In ASR systems the types of vocabularies can be classified as follows [4]

- Small vocabulary - tens of words
- Medium vocabulary - hundreds of words
- Large vocabulary - thousands of words
- Very-large vocabulary - tens of thousands of words
- Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word\

III. SPEECH RECOGNITION TECHNIQUES

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The speaker recognition system may be viewed as working in a four stage: Analysis, Feature extraction, Modeling, Testing/Matching techniques [4], [6].

A. Speech Analysis Techniques

Speech data contain different type of information such as vocal tract, excitation source and behaviour feature that shows a speaker identity. The speech analysis stage deals with segmenting speech signal into suitable frame size to be used for further analysis and extracting. The speech analysis technique done with following three techniques [4], [6]

1) *Segmentation analysis:* Speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information. Segmented analysis is used to extract vocal tract information of speaker recognition.

2) *Sub segmental analysis:* Speech is analyzed using the frame size and shift in range 3-5 ms. 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 by using the frame size and shift of 100-300 ms to extract speaker information mainly due to behavioral tract. These include word duration, intonation, speaker rate, accent etc.

B. Feature Extraction Techniques

Feature extraction step finds the set of parameters of utterances that have acoustic correlation with speech signals and these parameters are computed through processing of the acoustic waveform. These parameters are known as features. There are several methods for feature extraction such as Mel-Frequency Cepstral Coefficient (MFCC) [8], Linear Predictive Cepstral Coefficient (LPCC), Perceptual Linear Prediction (PLP), wavelet and RASTA-PLP (Relative Spectral Transform) Processing etc[5]. The most widely used feature extraction techniques are

1) *Linear Predictive Coding (LPC):* LPC has capability for speech compression, synthesis and as well as identification .LPC is spectral estimation technique because it provides an estimate of the poles of the vocal tract transfer function [7]. The LPC calculates a power spectrum of the signal [8]. It is predominant technique for determining the basic parameters of speech and provides precise estimation of speech parameters and computational model of speech. Speech sample can be approximated as a linear combination of past speech samples [9].

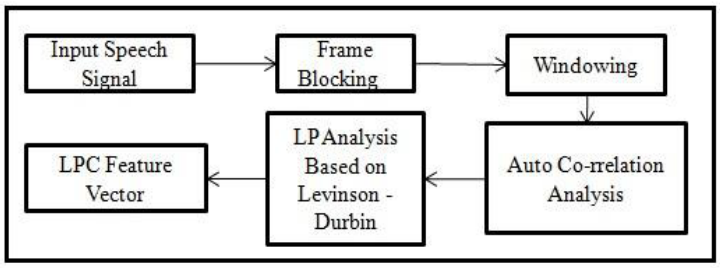

Fig.2 LPC Feature Extraction Process[10]

2) *Linear Predictive Cepstral Coefficients (LPCC):* When linear predictive coefficient is represented in cestrum domain, the obtained coefficients are linear predictive cepstral coefficients. Cestrum is obtained by taking inverse DFT of logarithm of the magnitude of the DFT of the speech signal [7].
3) *Mel- Frequency Cepstral Coefficients (MFCC's):* MFCC technique is based on the human peripheral auditory system [11]. Mel Frequency Cepstral Coefficients are based on the known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. The main purpose of the MFCC is to copy the behaviour of human ears [12].

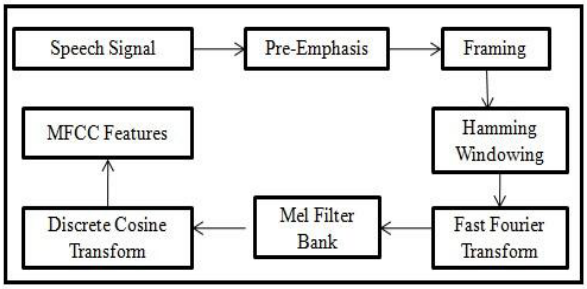

Fig.3 Block Diagram of MFCC Feature Extraction Techniques [10], [19]

4) *RASTA Filtering*: RASTA is short for RelAtive SpecTrAl. It is a technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the representations of the speech signals are band pass filtered in RASTA. Initially, it was just used to lessen the impact of noise in speech signal but know it is also used to directly enhance the signal [12]. The RASTA filter can be used either in the log spectral or cepstral domains [13].
5) *Zero Crossing with Peak Amplitudes (ZCPA):* This feature extraction technique is based on Human Auditory System. It uses zero-crossing interval to represent signal frequency information and amplitude value to represent intensity information, finally frequency information and amplitude information is combined to form the complete feature output [10].

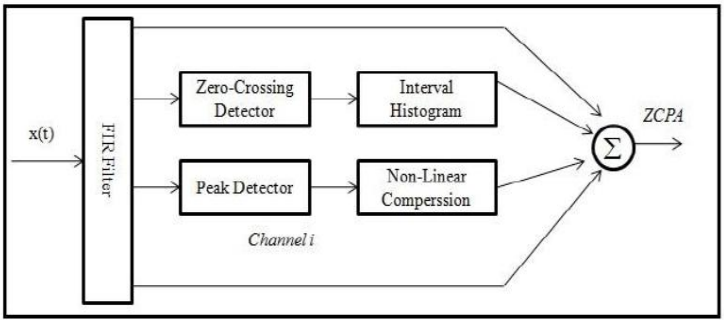

Fig.4 Principle diagram of ZCPA for Feature Extraction Techniques [10]

C. Modeling Techniques

The objective of modeling technique is to generate speaker models using speaker specific feature vector [6].

1) *Acoustic Phonetic Approach:* This approach uses knowledge of phonetics & linguistics to guide search process [15]. Acoustic phonetic approach for speech recognition is based on finding speech sound and providing appropriate labels these sounds. The basis of acoustic phonetic approach based on the fact that, there exist finite and exclusive phonemes in spoken language and these are broadly characterized by a set of acoustic properties that are demonstrated in the speech signal over time [9], [14]. This approach identifies individual phonemes, words, sentence structure and/or meaning [15]. Steps included in acoustic phonetic approach are as follows: The first step is the spectral analysis of speech which describes the broad acoustic properties of different phonetic units. The next step is segmentation and labeling the speech, that results in a phoneme lattice characterization of the speech. The last step is determination of string of words or a valid word from phonetic label sequences brought out by the segmentation to labeling. This approach has not been most extensively used in most commercial application [6], [9].

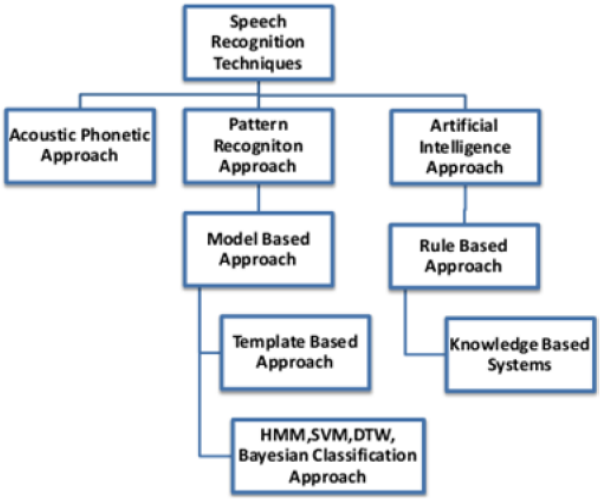

Fig 5 Speech Recognition Techniques [9], [14]

2) *Pattern Recognition Approach:* This method has two steps i.e. training of speech patterns and recognition of pattern by way of pattern comparison. In the parameter measurement phase (filter bank, LFC, DFT), a sequence of measurements is made on the input signal to define the "test pattern". The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern & reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase (dynamic time warping) [15]. The pattern-matching approach 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 [4], [6], [14], [15]. A speech pattern representation can be in the form of a speech template or a stochastic model and can be applied to a sound (smaller than a word), a word, or a phrase [4], [6].
3) *Template based Approach:* In template based approach, 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 [4, 16]. But it also has the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modelled by using many templates per word, which eventually becomes impractical [16]. 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 [4].
4) *Stochastic based Approach:* Stochastic modeling 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, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition [6], [17]. There exits many methods in this approach like HMM, SVM, DTW, VQ etc. The most popular stochastic approach today is hidden Markov modeling [9]. 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 [6], [17].
5) *Artificial Intelligence Approach:* The Artificial Intelligence approach [23] is a hybrid of the acoustic phonetic approach and pattern recognition approach. 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 [4], [16]. Acoustic phonetic knowledge is used to developed classification rules for speech sound where template based methods provide less insight about human speech processing, but these methods have been very productive in the design of a diversity of speech recognition system. This approach is not much successful as complexness in quantifying skilful knowledge. Integration of levels of human knowledge i.e. phonetics, lexical access, syntax and semantics, is the another problem of this approach. Artificial Neural Network method is more reliable method for this approach [9], [17].

D. Matching Techniques

Speech-recognition engines match a detected word to a known word using one of the following techniques.

1) *Whole-word matching:* The engine compares the incoming digital-audio signal against a prerecorded 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 hundred 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 [4], [16].

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 per word). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand [4], [16].

IV. APPLICATIONS OF SPEECH RECOGNITION

Some of the applications of speech recognition are [18]

- Data Entry Enhancements in an Electronic Patient Care Report (ePCR).
- Dictation.
- Command and Control.
- Telephony.
- Wearable.
- Medical/Disabilities.
- Embedded Applications.
- Agricultural application to get farmer queries.

V. CONCLUSION

In this paper, a comprehensive survey on speech recognition is provided. The fundamentals and classification of speech recognition system has also been discussed. The review of different approaches of speech recognition system has been presented.

REFERENCES

[1] V.Vaidhehi, Anusha J, Anand P, "The Comparative Study of Speech Recognition Models-Hidden Markov and Dynamic Time Wrapping Model", *International Research Journal of Innovative Engineering*, Volume1, Issue 3, pp. 22-28, March 2015

[2] Kimberlee A. Kemble, An Introduction to Speech Recognition, [online] Available: ftp://ftp.software.ibm.com/software/partners/comarketing/na/ss/we/WS_Voice_Server_White_Paper.pdf

[3] M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review", *International Journal of Computer Science and Information Security, Vol. 6, No. 3, pp. 181-205, 2009*

[4] Om Prakash Prabhakar, Navneet Kumar Sahu, "A Survey On: Voice Command Recognition Technique", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 5, pp. 576-585, May 2013

[5] Suman K. Saksamudre, P.P. Shrishrimal, R.R. Deshmukh,"A Review on Different Approaches for Speech Recognition System", *International Journal of Computer Applications, Volume 115 – No. 22, pp. 23-28, April 2015*

[6] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique*", International Journal of Computer Applications*, Volume 10– No.3, pp. 16-24, November 2010

[7] Gaganpreet Kaur, Dr. Dheerendra Singh, " A Survey on Speech Recognition Algorithms", *International Journal of Emerging Research in Management &Technology, Volume-4, Issue-,pp. 289-292, May 2015*

[8] Namrata Dave, " Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", *International Journal for Advance Research in Engineering and Technology*, Volume 1, Issue VI, pp. 1-5, July 2013

[9] Nidhi Desai, Prof.Kinnal Dhameliya, Prof.Vijayendra Desai, "Feature Extraction and Classification Techniques for Speech Recognition: A Review", *International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Issue 12, pp. 367-371, December 2013

[10] Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 3, Issue 12, pp. 18006-18016, December 2014

[11] Rashmi C R, "Review of Algorithms and Applications in Speech Recognition System", *International Journal of Computer Science and Information Technologies*, Vol. 5 (4), 2014, pp. 5258-5262

[12] Shreya Narang, Ms. Divya Gupta, "Speech Feature Extraction Techniques: A Review", *International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 3, March 2015, pp. 107 – 114*

[13] *Urmila Shrawankar, Dr. Vilas Thakare, "* TECHNIQUES FOR FEATURE EXTRACTION IN SPEECH RECOGNITION SYSTEM : A COMPARATIVE STUDY", International Journal Of Computer Applications In Engineering, Technology and Sciences (IJCAETS),2010, pp 412-418

[14] Shanthi Therese S.,Chelpa Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition*", International Journal of Scientific Engineering and Technology*, Volume No.2, Issue No.6, pp : 479-484, 1 June 2013

[15] Preeti Saini, Parneet Kaur, "Automatic Speech Recognition: A Review", *International Journal of Engineering Trends and Technology- Volume4Issue2, pp.132-136, 2013*

[16] Muhirwe Jackson, "AUTOMATIC SPEECH RECOGNITION: HUMAN COMPUTER INTERFACE FOR KINYARWANDA LANGUAGE", M. Sc., Makerere University, August 2005

[17] Pradeep Kumar Jaisal, Pankaj Kumar Mishra, "A Review of Speech Pattern Recognition: Survey", IJCST Vol. 3, Iss ue 1, pp. 709-713, Jan. - March 2012

[18] Dr.E.Chandra, A.Akila, "An Overview of Speech Recognition and Speech Synthesis Algorithms", Int.J.Computer Technology & Applications,Vol 3 (4), pp. 1426-1430, July-August 2012

[19] Dr E.Chandra, K.Manikandan, M. Sivasankar, "A Proportional Study on Feature Extraction Method in Automatic Speech Recognition System", *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING Vol. 2, Issue 1, pp. 772-775, January 2014*

