Speech Processing: A Review

[1]Sunita Dixit, [2]Dr. MD Yusuf Mulge
[1]Research scholar, Pacific university, Udaipur
[2]Principal, PDM College of Engineering for Women, Bahadurgarh

Abstract: Speech is a complex signal that is characterized by varying distributions of energy in time as well as in frequency, depending on the specific sound that is being produced. The aim of digital speech processing is to take advantage of digital computing techniques to process the speech signal for increased understanding, improved communication, and increased efficiency and productivity associated with speech activities. In this paper, we have reviewed various phases of speech processing and presented some difficulties that arises while speech recognition.

Keywords: speech processing, speech recognition, feature extraction, verification.

I. INTRODUCTION

Speech signals are composed of a sequence of sounds and the sequence of sounds are produced as a result of acoustical excitation of the vocal tract when air is expelled from the lungs. There are various ways to categorize speech sounds. Speech sounds based on different sources to the vocal tract[1]. Speech sounds generated with a periodic glottal source are termed Voiced. Voiced speech is produced when the vocal cords play an active role (i.e. vibrates) in the production of the sound. Examples /a/, /e/, /i/. Likewise, sounds not generated are called unvoiced. Unvoiced sounds are produced when vocal cords are inactive. Examples /s/,/f/. Other classes are Nasal Sounds and Plosives. Nasal sounds are the one in which sound gets radiated from nostrils and lips. Examples include /m/, /n/, /ing/. Plosive sounds are those characterized by complete closure /constriction towards front of the vocal tract. Examples include /p/,/t/. Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as following.

2.1 Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses)[2]. Isolated Utterance might be a better name for this class.

2.2 Connected Words

Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be run-together with a minimal pause between them.

2.3 Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. (Basically, it's computer dictation). Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

2.4 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 features.

II. SPEECH PROCESSING

The aim of digital speech processing is to take advantage of digital computing techniques to process the speech signal for increased understanding, improved communication, and increased efficiency and productivity associated with speech activities[3]. There are three main phases of speech processing:

- speech analysis,
- speech recognition,
- speech coding.

Figure 1 shows speech processing process in detail representing three main phases. It also depicts main phases associated with recognition process.

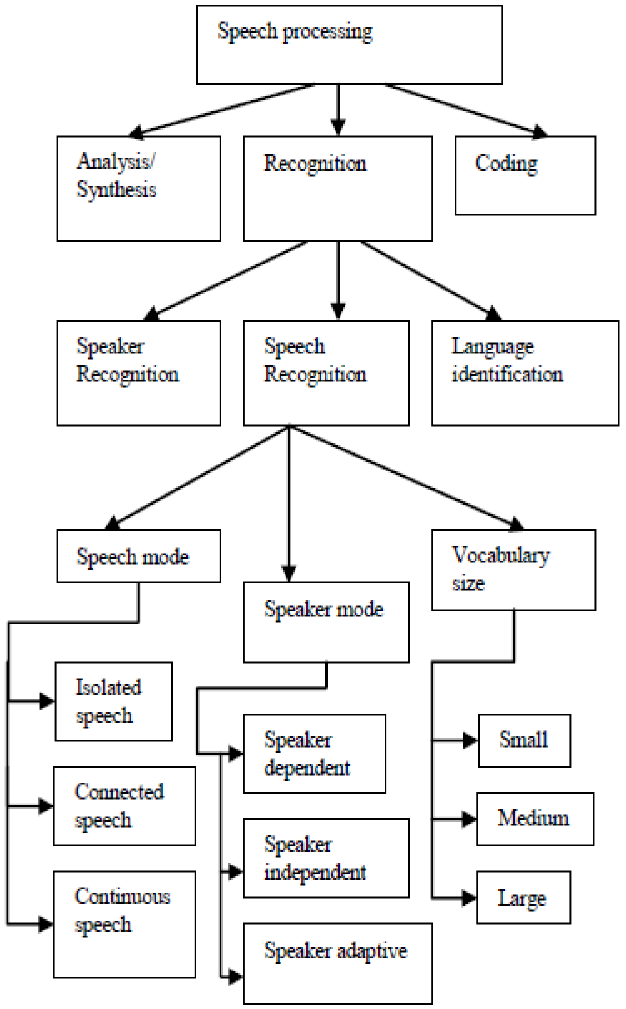

Fig.1 speech processing

A. Speech analysis

The speech analysis stage deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting [7]. The speech analysis technique done with following three techniques.

- *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.

- *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. [8].

- *Supra Segmental Analysis*

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

B. Speech recognition

Speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). Some SR systems use "speaker-independent speech recognition" while others use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine-tune the recognition of that person's speech, resulting in more accurate transcription.

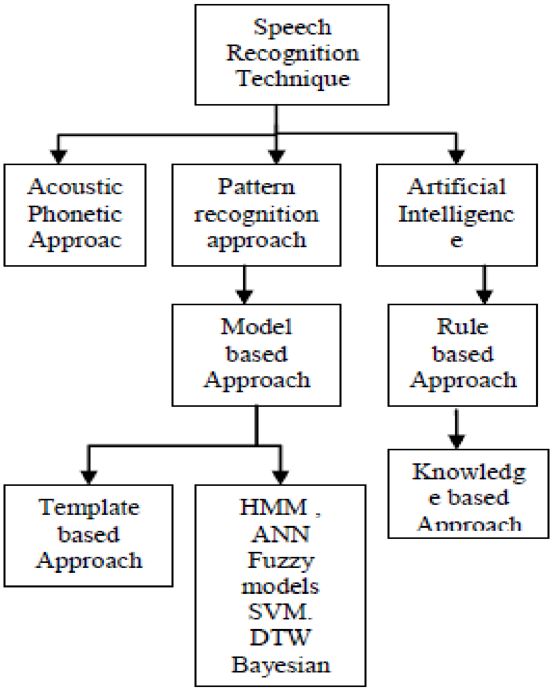
Speech Recognition Technique
Acoustic Phonetic Approac
Pattern recognition approach
Artificial Intelligenc e
Model based Approach
Rule based Approach
Knowledg e based Approach
Template based Approach
HMM , ANN Fuzzy models SVM. DTW Bayesian

Fig2 speech recognition techniques

Speech recognition techniques can be broadly classified in three categories:

- *Acoustic Phonetic Approach*

In this speech recognition algorithm, the system tries to decode the speech signal in a sequential manner based on the observed acoustic features of the speech waveform and the known relations between acoustic features and phonetic symbols.

- *Pattern Recognition Approach*

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.

- *Artificial Intelligence Approach*

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. 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. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand.

C. Speech Coding

One of the most important applications of digital speech processing is speech coding, which is concerned with efficient and reliable communication between people who may be separated by geographical distance or by time. The former forms the basis of modern telephony, which enables people to converse regardless of their locations and the latter forms the basis for applications like “voice mail” which let people create and retrieve verbal messages at arbitrary times. Speech coding enables both telephony and voice messaging by converting the speech signal into a digital format suitable for either transmission or storage. Relevant issues in speech coding are conservation of bandwidth (the rate of the voice coder), voice quality requirements, processing and transmission delay, and processing power, as well as techniques for privacy and secure communication.

III. DIFFICULTIES IN SPEECH RECOGNITION

A major obstacle in obtaining high-accuracy recognition is the large variability in the speech signal characteristic. The three components of variability are: linguistic variability, speaker variability, and channel variability. Linguistic variability includes the effects of phonetics, phonology, syntax, semantics, and discourse on the speech signal. Speaker variability includes intra- and inter-speaker variability and the effects of co-articulation. Channel variability includes the effects of background noise and the transmission channels (e.g., microphone, telephone, and reverberation). The above-mentioned variabilities sometimes interfere with the intended message and the problem must be unraveled by the recognition process.

Robustness against speech variation is one of the most important issues in speech and speaker recognition. There are many causes of speech variation. The main causes of speech variation can be classified based on whether they originate in the speaking and recording environment. Another set of speech recognition problems involves the type of hardware used to actually input the sound, because the results can have a large impact in how the software will interpret the speech. There also is the problem of not knowing the context of the words being spoken, which can lead to text that has no punctuation or inaccurate spellings. Some other difficulties in speech recognition include:

- Homophones ambiguity
- Word boundary ambiguity
- Regional and social dialects
- Speaking style
- Noise
- Out of vocabulary

IV. CONCLUSION

In this paper, we have reviewed, classified various phases of speech processing. Most important phase of speech processing is speech recognition which involves three main approaches-acoustic phonetic approach, pattern matching approach, & artificial intelligence approach. The main goal of speech recognition is to get efficient ways for humans to communicate with computers. The problem of automatically recognizing speech with the help of a computer is a difficult problem, and the reason for this is the complexity of the human language. The paper also highlights various difficulties in speech recognition.

V. REFERENCES

[1] Simon Kinga and Joe Frankel, Recognition, “Speech production knowledge in automatic speech recognition”, Journal of Acoustic Society of America, Oct 2006.

[2] H.Hermansky, “Perceptual Linear Predictive Analysis of Speech,” The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990

[3] D.O. Shaughnessy, Speech Communication: Human and Machine. Second Edition India: University Press (India) Private Limited, 2001 [12] Fang Zheng , Guoliang Zhang , and Zhanjiang Song “Comparison of Different Implementations of MFCC”, The journal of Computer Science & Technology, pp. 582-589, Sept. 2001

[4] Hossan, M.A. “A Novel Approach for MFCC Feature Extraction”, 4th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, Dec 2010

[4] Oh-Wook Known, Kwokleung Chan, Te-Won Lee, “Speech Feature Analysis Using Variational Bayesian PCA”, in IEEE Signal Processing letters, Vol. 10, pp.137 – 140, May 2003

[5] Santosh Gaikwad, Bharti Gawali, Pravin Yannawar, Suresh Mehrotra, “Feature Extraction Using Fusion MFCC For Continuous Marathi Speech Recognition”, in IEEE conference (INDICON), pp 1-5, Dec.2011

[6] Anup Kumar Paul, Dipankar Das and Md. Mustafa Kamal, “Bangla Speech Recognition System using LPC and ANN,” Seventh International Conference on Advances in Pattern Recognition, pp. 171 – 174, Feb.2009

[7] H.Hermansky, B. A. Hanson, H. Wakita, "Perceptually based Linear Predictive Analysis of Speech," Proc. IEEE Int. Conf. on Acoustic, speech, and Signal processing, pp. 509-512, Aug.1985

[8] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based Processing in Automatic Speech Recognition," Proc. IEEE Int. Conf. on Acoustic, speech, and Signal processing, pp. 1971-1974, Apr.1986.

[9] G. R. Doddington, "Speaker recognition—Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651–1664, Nov. 1985.

[10] A. L. Higgins and R. E. Wohlford, "A new method of text independent speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1986, pp. 869–872.

[11] A. Higgins, L. Bhaler, and J. Porter, "Voice identification using nearest neighbor distance measure," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis, MN, 1993, pp. 375–378.