INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

A Review on Speech Recognition with Filters

Deepak[1], Vikas Mittal[2]

M.M. Engineering College, M.M. Engineering College
Mullana (Ambala), India, 133203
[1]deepak.evergreen@gmail.com

M.M. Engineering College, M.M. Engineering College
Mullana (Ambala), India, 133203
[2]vikasmittal2k7@rediffmail.com

***Abstract*: -** *Speech recognition is a popular topic in today's life because of its numerous applications. For example, consider the applications in the mobile phone in which instead of typing the name of the person who user want to call, the user can just directly speak the name of person to the mobile phone and the mobile phone will automatically call that person. The Speech is most prominent & primary mode of Communication among of human being. However, speech is a random phenomenon and thus its recognition is a very challenging task. This paper gives an overview of major technological aspects also gives overview technique developed in each stage of speech recognition.*

***Key words*:** *Speech recognition, feature extraction, WER, Acoustic phonetic, Dynamic Time Wrapping.*

1. INTRODUCTION

Speech recognition is a technique that enables a device to recognize and understand spoken words, by digitizing the sound and matching its pattern against the stored patterns. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. A speech recognition system consists of a microphone for the person to speak into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans.

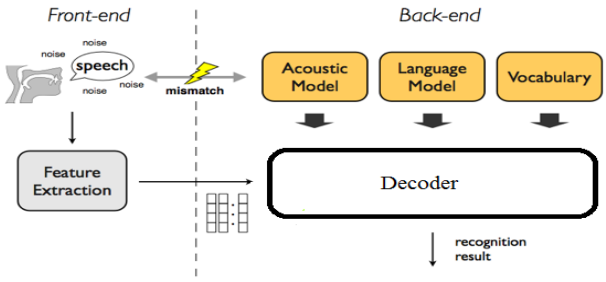

Figure 1: Basic Architecture of Speech Processing [8]

The basic architecture of speech processing can be divide into two parts: Front end and Back end. The front end basically comprises of the speech or signal generally intermixed with noise and the feature extraction unit which plays a vital role in the speech recognition. The process of speech recognition initialize when the speaker starts speaking and generates sound waves. The sound waves are then captured by the microphone or the sound card. Originally the speech is in the form of wave which later on converted into electrical energy by microphone. These signals are then converted to the discrete signals which are assumed to contain only the relevant information about given utterance that is important for its correct recognition.

The goal of speech recognition area is to developed technique and system to developed for speech input to machine based on major advanced in statically modeling of speech ,automatic speech recognition today find widespread application in task that require human machine interface such as automatic call processing [2].

1.1 Topology of Speech Recognition Systems

- Speaker Dependent: - systems that require a user to train the system according to his or her voice.
- Speaker Independent: - systems that do not require a user to train the system i.e. they are developed to operate for any speaker.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

- Isolated word recognizers: - accept one word at a time. These recognition systems allow us to speak naturally continuous.
- Spontaneous recognition systems allow us to speak spontaneously [3].

1.2 Types of Speech Recognition

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 [1,7](usually doing processing during the pauses). Isolated Utterance might be a better name for this class .

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.

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 [1,7].

1.3 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. direction machines)[6]. In ASR systems the types of vocabularies can be classified as follows.

- Small vocabulary - ten 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

Apart from the above characteristics, the environment variability, channel variability, speaker style, sex, age, speed of speech also make the speech recognition (SR) system more complex. But the efficient SR systems must cope with the variability in the signal[6].

2. SPEECH RECOGNITION TECHNIQUES

The goal of speech recognition is to make a machine 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 goal of automatic speaker recognition is to analyze, extract characterize and recognize information about the speaker identity. The speaker recognisation system may be viewed as working in a four stages [4]:

- Analysis
- Feature extraction
- Modeling
- Testing

2.1 Speech analysis

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. This uniqueness is embedded in the speech signal during speech production and can be used for speaker used for speaker recognition. The behavioral tracts as to how the vocal tract and excitation source are controlled during speech production are also unique for each user. The information about behavioral tracts is also embedded in the speech signal and can be used for speaker recognition. The information about the behavior feature also embedded in signal and that can be used for speaker recognition. The speech analysis deals with stages with suitable frame size for segmenting speech signal for further analysis and extracting. The speech analysis is technique done with following three techniques [4]. Segmentation analysis, Sub-segmental analysis and Supra-segmental analysis.

2.2 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.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

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 speech feature extraction in a categorization problem is about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. As we know from fundamental formation of speaker identification and verification system that the number of training and test vector needed for the classification problem grows with the dimension of the given input so we need feature extraction of speech signal. The purpose of feature extraction stage is to extract the speaker-specific information in the form of feature vectors. The feature vectors represent the speaker-specific information due to one or more of the following: Vocal tract, excitation source and behavioral tracts. A good feature set should have representation due to all of the components of speaker information. Just as a good feature set is required for a speaker, it is necessary to understand the different feature extraction techniques [4].

2.3 Speaker Modeling Technique

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided into two classifications: speaker recognition and speaker identification. The speaker identification technique automatically identify who is speaking on basis of individual information integrated in speech signal The speaker recognition is also divided into two parts that means speaker dependant and speaker independent. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message, on the one hand. On the other, in case of speaker recognition, machine should extract speaker characteristics in the acoustic signal. The main aim of speaker identification is comparing a speech signal from an unknown speaker to a database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be divided into two methods, text-dependent and text-independent methods. In text-dependent method, the speaker says 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. Following are the modeling which can be used in speech recognition process: The acoustic-phonetic approach, Pattern Recognition Approach, Template based approaches, Dynamic Time Warping (DTW), The Artificial Intelligence Approach [4].

2.4 Testing

Testing can be done by using various models or algorithms by implementing them either with the help of software or by implementing hardware. Implementation can be done by creating database of voice samples of different users. Samples can be collected with the help of microphone mounted to computer or laptop. Noise can be removed with the help of various algorithms or by implementing various filters. Test sample can be matched with the samples stored in database and required results can be obtained [4].

3. HOW TO APPROACH

There are mainly three ways to approach to the speech recognition process. This approach makes the speech recognition process much easier. These approaches are as follows:

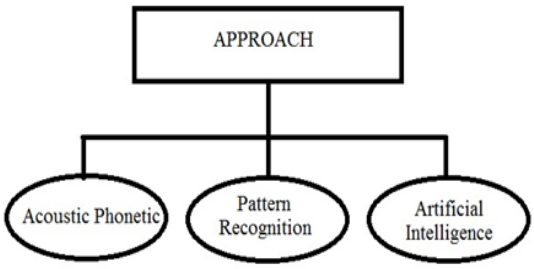

Figure 2: Approach to Speech recognition

3.1) Acoustic Phonetic: It is a subfield of phonetics which deals with acoustic aspects of speech sounds. Acoustic phonetics investigates properties like the mean squared amplitude of a waveform, its duration, its fundamental frequency, or other properties of its frequency spectrum. The earlier method to approach speech recognition was based on finding speech sounds and providing appropriate labels to these sounds. Even though, the acoustic properties of phonetic units are highly variable. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling. The acoustic phonetic approach has not been widely used in most commercial applications [5].

3.2) Pattern Recognition: Pattern recognition as the name suggests recognize the pattern of the speech sound. This can be done by two methods i.e by pattern comparison or by pattern training[3].To check the parameter of the input signal various measurement is done on the input signal which defines 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[3].

3.3) Artificial Intelligence approach: It 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. 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. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics[1].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[5]. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported.

4. FILTERS ASSOCIATED TO SPEECH RECOGNITION

As there are various filters proposed by many researchers, here we are going to discuss two most promising filters for the purpose of speech recognition. Namely these are Wiener filter and kalman filter. These filters provide an ease for storing voice as they help in reduction in noise.

4.1) Wiener Filter: The Wiener filter is a popular technique that has been used in many signal recognition methods. The basic principle of the Wiener filter is to obtain an estimate of the clean signal from that corrupted by additive noise. This estimate is obtained by minimizing the Mean Square Error (MSE) between the desired signal and estimated signal [9]. The principle of FIR Wiener filter is shown in Figure 3 below.

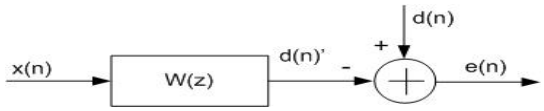

Figure 3: Wiener filter for speech recognition purposes [10]

It is used to estimate the desired signal d(n) from the observation process x(n) to get the estimated signal d(n). It is assumed that d(n) and x(n) are correlated and jointly wide-sense stationary. The Wiener filter obtains a least squares estimate of d(n)' under stationary assumptions of speech and noise. The construction of the Wiener filter requires an estimate of the power spectrum of the clean speech and the noise[10].

4.2) Kalman Filter: Kalman filter is an adaptive least square error filter that provides an efficient computational recursive solution for estimating a signal in presence of Gaussian noises. It is an algorithm which makes optimal use of imprecise data on a linear (or nearly linear) system with Gaussian errors to continuously update the best estimate of the system's current state. This method is best suitable for reduction of white noise to comply with Kalman assumption. In deriving Kalman equations it is normally assumed that the process noise (the additive noise that is observed in the observation vector) is uncorrelated and has a normal distribution. This assumption leads to whiteness character of this noise.

5. PERFORMANCE CALCULATION

Reliability of any system can be checked or calculated when its outcome is known. Outcome of a system depends upon its various factors. For a speech recognition system, performance is specified in terms

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

of accuracy and speed [5]. Accuracy is measured and rated with the word error rate (WER) and speed is measured as real time function. Word error rate (WER) is a common metric of the performance of a speech recognition or machine translation system. The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. Word error rate can then be computed as:

where, S is the no. of substitutions, D is the no. of deletions, I is the no. of insertion, N is the no. 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 [6],

Single word error rate (SWER) and command success rate (CSR) are some other performance parameter on which a speech recognition system relies [3]. When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead [5].

6. CONCLUSION

Since speech is the only best method of communication for human being and humans do a daily activity of speech recognition, hence speech processing has always remained an interesting topic for researchers. In this paper, the fundamentals of speech recognition are discussed. The various approaches available for developing an speech recognition system are clearly explained. Various performance parameters are also discussed which is a very vital part for any speech recognition system. We hope this paper brings a basic understanding to the researchers and motivate them to explore the community of speech recognition.

REFERENCES

[1] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and information Security, vol. 6, no. 3, pp.181-205, 2009.

[2] R.Klevansand R.Rodman, "voice Recognition, Artech House, Boston, London 1997.

[3] Preeti Saini and Parneet Kaur," Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology (IJETT)- Volume4Issue2- 2013.

[4] Sanjib Das, "Speech Recognition Technique: A Review", International Journal of Engineering Research and Applications-Vol.2, Issue 3, , pp.2071-2087,May-Jun 2012.

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

[6] Vimala.C and V.Radha," A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal (WCSIT)- Vol. 2, No. 1, 1-7, 2012.

[7] L. Deng and X. Huang (2004), "Challenges in adopting speech recognition", Communications of the ACM, 47(1), pp 69-75.

[8] ESAT speech group website. [online]. Available: http://www.esat.kuleuven.be/psi/spraak/theses/08-09-en/MDT.php

[9] M.A Abd El-Fattah et al," Speech enhancement using an adaptive wiener filtering approach", progress in electromagnetics research M, vol. 4, 167-184, 2008.

[10] Deepak and Vikas mittal," Speech Recognition using FIR Wiener Filter", International Journal of Application or Innovation in Engineering & Management (IJAIEM)- Volume 2, Issue 5, May 2013.