A Review on Speech Recognition

Pappu Mandal[1] Christi D'Souza[2] Ajay Pal[3] Shiwani Gupta[4]
[1,2,3]Research Student [4]Assistant Professor
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]Thakur College of Engineering & Technology

Abstract— Speech recognition is the new emerging technology in the field of computer and artificial intelligence. It has changed the way we communicate with computer and other intelligent devices of same calibre like smart phones. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper is concludes with the decision on feature direction for developing technique in human computer interface system.

Key words: Speech Recognition, Phonetics, Feature extraction, performance evaluation, HMM
General Terms: Modeling technique, speech processing, signal processing, Pattern Recognition

I. INTRODUCTION

Speech recognition is the translation of spoken words into text. It is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Speech Recognition is technology that can translate spoken words into text. Speech is a form of communication we learn early and practice often, so the use of speech recognition software can simplify computer interfaces and make computers accessible to users unable to key text using a standard keyboard. However, computer-based speech recognition is more difficult to achieve than one might at first assume. The speech recognition process is statistical in nature and is based on Hidden Markov Models (HMMs). The HMM is first trained using speech data for which the associated text is known. Subsequently, the trained HMM is used to "decode" new speech data into text.

A. Types of Speech:

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 based on the fact that one of the difficulties of ASR is the ability to determine when a speaker starts and finishes an utterance [11].

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 doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

2) Connected Words:
Connect 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.

3) Continuous Speech:
Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation [2][3].

4) Spontaneous Speech:
There appears to be a variety of definitions for what spontaneous speech actually is. 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 such as words being run together, "ums" and "ahs", and even slight stutters [2].

5) Voice Verification/Identification:
Some ASR systems have the ability to identify specific users. This document doesn't cover verification or security systems.

B. Speech Applications:

Speech recognition applications may be classified into three categories: dictation systems, navigational or transactional systems, and multimedia indexing systems. Each category of applications has a different tolerance for speech recognition errors. Advances in technology are making significant progress toward the goal of any individual being able to speak naturally to a computer on any topic and to be understood accurately [1].

1) Dictation Applications:
Such applications are those in which the words spoken by a user are transcribed directly into written text. Such applications are used to create text such as personal letters, business correspondence, or e-mail messages. Usually, the user has to be very explicit, specifying all punctuation and capitalization in the dictation. Dictation applications often combine mouse and keyboard input with spoken input. Using speech to create text can still be a challenging experience since users have a hard time getting used to the process of dictating. Best results are achieved when the user speaks clearly, enunciates each syllable properly, and has organized the content mentally before starting. As the user speaks, the text appears on the screen and is available for correction. Correction can take place either with traditional methods such as a mouse and keyboard, or with speech [1].

2) Transactional Applications.
Speech is used in transactional applications to navigate around the application or to conduct a transaction. For example, speech can be used to purchase stock, reserve an airline itinerary, or transfer bank account balances. It can also be used to follow links on the web or move from application to application on one's desktop. Most often, but not exclusively, this category of speech applications

involves the use of a telephone. The user speaks into a phone, the signal is interpreted by a computer (not the phone), and an appropriate response is produced. A custom, application-specific vocabulary is usually used; this means that the system can only "hear" the words in the vocabulary. This implies that the user can only speak what the system can "hear." These applications require careful attention to what the system says to the user since these prompts are the only way to cue the user as to which words can be used for a successful outcome [4].

3) Multimedia Indexing Applications.

In multimedia indexing applications, speech is used to transcribe words from an audio file into text. The audio may be part of a video. Subsequently, information retrieval techniques are applied on the transcript to create an index with time offsets into the audio. This enables a user to search a collection of audio/video documents using text keywords. Retrieval of unstructured multimedia documents is a challenge; retrieval using keyword search based on speech recognition is a big step toward addressing this challenge. It is important to have realistic expectations with respect to retrieval performance when speech recognition is used. The user interface design is typically guided by the "search the speech, browse the video" metaphor where the primary search interface is through textual keywords, and browsing of the video is through video segmentation techniques. In general, it has been observed that the accuracy of the top-ranking search results is more important than finding every relevant match in the audio. So, speech indexing systems often bias their ranking to reflect this. Since the user does not directly interact with the indexing system using speech input, standard search engine user interfaces are seamlessly applicable to speech indexing interfaces [3].

II. WORKING OF SR SYSTEM

Speech recognition basically means talking to a computer, having it recognize what we are saying, and lastly, doing this in real time. This process fundamentally functions as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech (fig.1). The elements of the pipeline are:

A. Transform the PCM Digital Audio into a Better Acoustic Representation:

The input to speech recognizer is in the form of a stream of amplitudes, sampled at about 16,000 times per second. But audio in this form is not useful for the recognizer. Hence, Fast-Fourier transformations are used to produce graphs of frequency components describing the sound heard for 1/100 of a second [10].

B. Unit Matching System:

It provides likelihoods of a match of all sequences of speech recognition units to the input speech. These units may be phones, diaphones, syllables or derivative units such as fenones and acoustic units. They may also be whole word units or units corresponding to group of 2 or more words. Each such unit is characterized by some HMM whose parameters are estimated through a training set of speech data.

C. Lexical Decoding:

constraints the unit matching system to follow only those search paths sequences whose speech units are present in a word dictionary.

D. Apply a "Grammar":

so the speech recognizer knows what phonemes to expect. This further places constraints on the search sequence of unit matching system. A grammar could be anything from a context-free grammar to full-blown English.

III. ALGORITHM USED FOR SPEECH RECOGNITION

Both acoustic modelling and language modelling are important parts of modern statistically-based speech recognition algorithms. Hidden Markov models (HMMs) are widely used in many systems. Language modelling has many other applications such as smart keyboard and document classification.

Hidden Markov models- Modern general-purpose speech recognition systems are based on Hidden Markov Models[1]. These are statistical models that output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short timescales (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for much stochastic purposes. Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of n-dimensional real-valued vectors (with n being a small integer, such as 10), outputting one of these every10 milliseconds. The vectors would consist of kestrel coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a Fig 1 shows the general speech recognition system. Each word, or (for more general speech recognition systems), mixture of diagonal covariance Gaussians, which will give likelihood for each observed vector [7]. Each phoneme will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

A. Mel Frequency Cepstral Coefficients:

MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency Cepstral these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as

$$m=2595 \log_{10}(1+f/700) \quad (1)$$

Mel scale and normal frequency scale is referenced by defining the pitch of 1000 Mel to a 1000 Hz tones, 40 db above the listener's threshold. Mel frequency are equally spaced on the Mel scale and are applied to linear space filters below 1000 Hz to linearized the Mel scale values and logarithmically spaced filter above 1000 Hz to find the log power of Mel scaled signal [13] [14]. Mel frequency wrapping is the better representation of voice. Voice features are represented in MFCC by dividing the voice signal into frames and windowing them then taking the Fourier transform of a windowing signal. Mel scale frequencies are obtained by applying the Mel filter or triangular band pass filter to the transformed signal. Finally transformation to the discrete form by applying DCT presents the Mel Cepstral Coefficients as acoustic features of human voice.

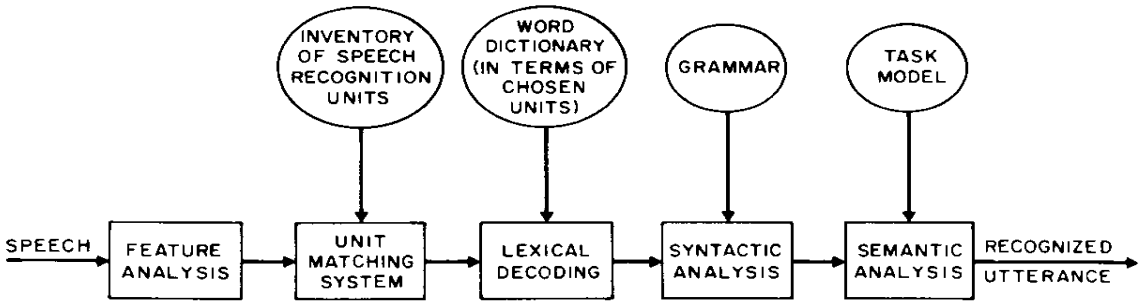

Fig. 1: Block diagram of a speech recognizer

IV. FEATURE SCOPE:

With the help of above review we can design and implement speech recognition and rectification system for articulatory handicapped people which will be a noble work for society. And hence we can reduce the speech communication problems faced by articulatory handicapped people in their day to day life.

V. CONCLUSION

In this review, we have discussed the technique developed in each stage of speech recognition system. . MFCC is used to extract the voice features from the voice sample. And HMM is used to recognize the speaker on the basis of extracted features. Through this review it is found that MFCC is used widely for feature extraction of speech and HMM is best among all modeling technique.

VI. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

REFERENCES

[1] Rabiner, L. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." Proceedings of IEEE 77, no. 2 (1989):257-286.
[2] Schmandt, C. Voice Communications with Computers. New York*:* Van Nostrand Reinhold, 1994.
[3] Wactlar, H., et al. "Lessons Learned from Building a Terabyte Digital Video Library." IEEE Computer (1999): 66-73.
[4] Stefan Eickeler, K. Biatov, Martha Larson, J. Kohler, Two Novel Applications of Speech
[5] Recognition Methods for Robust Spoken Document Retrieval.
[6] Karat, C., et al. "Patterns of Entry and Correction in Large Vocabulary Continuous
[7] Speech Recognition Systems." Proceedings of CHI '99: Human Factors in Computing Systems, *(1999):* 568-575.
[8] Brian Delaney, Tajana Simunic, Nikil Jayant, Energy Aware Distributed Speech Recognition for Wireless Mobile Devices
[9] http://www.zachary.com/s/xvoice
[10] Surabhi Bansal, Ruchi Bahety "Speech recognition system" From https://cseweb.ucsd.edu/classes/fa06/cse237a.html
[11] Speech Recognition Types From http://tldp.org/HOWTO/Speech-Recognition-HOWTO/introduction.html
[12] Speech recognition using hmm with mfcc- an analysis using frequency specral decomposion technique Source:Signal & Image Processing : An International Journal(SIPIJ) Vol.1, No.2, December 2010
[13] Anjali Bala, Abhijeet Kumar and Nidhika Birla, "Voice Command Recognition System Based On MFCC and DTW" Anjali Bala et al. / International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342
[14] http://en.wikipedia.org/wiki/Mel_scale

