A Comparative Study on Speech Recognition Approaches and Models

K Naga Abhishek Reddy, Parul Agrawal, Poonam Singh, Prerna Singh, Latha N.R

(Assistant Professor) Computer Science and Engineering, BMS College of Engineering, Bangalore, Karnataka, India

Abstract: *This paper is a study on various speech recognition and speech to text conversion approaches. Speech recognition applications are finding its importance today, and various interactive speech aware applications are available in the market. Speech recognition has become one of the widely used technologies, as it offers great opportunity to interact and communicate with automated machines in various parts of life, right from walk y talkies to mobile phones, microphones to earphones, we use speech or voice inputs. Recent researches have revealed that speech recognition has various issues that affect the decoding of speech. In order to overcome these issues, different models are developed by the researchers. This study is an effort to compare different approaches and models applied for speech recognition based on their type, advantages disadvantages and how voice input is taken into the system.*

Keywords: *Speech recognition, comparisons, different approaches of speech recognition***.**

I. Introduction

Speech recognition can be stated as a process that makes a computer recognize and respond to the sounds produced in human language. The need for speech recognition is due to the increased evolution of technology, where data is obtained in an ultra-speed, data transferred in ultra-speed through fibour optics and such systems. And the best way to communicate is through common language or rather only by speech. Hence the need of speech recognition is important to evolve the data transfer and data obtained in the form of an end-to-end talking, particularly speech recognition and speech exchange. However a speech recognition stores limited words in its vocabulary, it may be used to generate various results by storing its vocabulary with required data. Therefore, a comparative study on various models has been carried out in this paper which concludes which model or method to be used for speech recognition.

II. Literature Survey

Reference [1] discusses two methods that are feasible when realizing real-time speech-to-text transfer: speech recognition and computer assisted note taking (CAN). In Automatic speech recognition (ASR), the speaker has to train the speech recognition system in advance with the voice and speaking characteristics. This process is based on a match of physical parameters of the actual speech signal with a representation which was generated on the basis of a general phonetic model of language and the phonetic and voice data from the individual. With computer-assisted note taking (CAN), a person writes into an ordinary computer what a speaker says. However even professional writing speed is not sufficient to write down every word of a speech and nor this system is efficient or feasible.

Reference [2] discusses about Template-Based Approach that has been used for speech-to-text conversion. The output from the signal processing module of a speech recognizer is a sequence of feature vectors. One first builds a collection of reference templates, each itself a sequence of feature vectors that represents a unit (usually a whole word) of speech to be recognized. Then, the feature vector corresponding to the current utterance is compared with each reference vector in turn, via some distance measure. The template-based approach does not seem extensible to IWR and CSR when the dictionary and language model are large. In these cases, the stochastic approach based on maximum likelihood is applied.

Reference [3] discusses about an electronic device called an Analog- to-digital converter (ADC) that has been used to convert the Analog speech waveform to a series of digital signals. These signals are stored as a template and carry the time-based characteristics for each word spoken. The spoken words are identified and then the pattern-matching algorithm (HMM) is used to search in all the grammatically possible word. Automatic Speech Recognition (ASR) is the process by which a computer maps an acoustic speech signal to text (known as ASR technology). Speech-to-text recognition can be classified as first individual character recognition in which characters uttered in isolation are recognized, second an isolated word recognition, in which words uttered in isolation are recognized, and third a continuous speech recognition i.e. connected word recognition. The objective of speech recognition is to recognize the message being spoken.

Reference [4] shows the existing statistical models for speech conversion. One of them is Acoustic Model. It captures the characteristics of the basic recognition units. This conversion is done at the phoneme level for which large vocabulary continuous speech conversion systems is used (Ex: Hidden Markov models and neural networks).

The other model is the Language Model (LM). One of the major objectives of language model is to convey or transmit the behaviour of the language. It is due to the fact that it intends to forecast the existence of the specific word sequences within the target speech. It uses statistical LM toolkit.

Lexicon model provides the pronunciation of the words within the target speech, which has to be recognized. It is based on two parameters, i.e., whole-word access, and decomposition of entire speech into small chunks. This process eventually results in appropriate conversion of the speech. Hidden Markov Models (HMM) is a statistical tool, which is used for the modelling of data. This model has reduced the issues of speech classification (continuous or discrete) which was one of the core issues.

Reference [5] uses a technique of transforming the PCM digital audio into a better acoustic. Fast-Fourier transformation is used. Unit matching system -provides likelihoods of a match of all sequences of speech recognition units to the input speech. Lexical Decoding -constraints the unit matching system to follow only those search paths sequences whose speech units are present in a word dictionary. Apply a "grammar"- so the speech recognizer knows what phonemes to expect. Figure out which phonemes are spoken. A speech recognizer works by hypothesizing a number of different "states" at once.

Reference [6] discusses the conversion of PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech. The PCM digital audio is transformed into a better acoustic representation. Grammar is applied so the speech recognizer software knows what phonemes to expect. The software is supposed to figure out which phonemes are spoken which are then converted into words.

The first element of the pipeline converts digital audio coming from the sound card by sampling it. The PCM digital audio is transformed into the "frequency domain." This is done using a windowed fast-Fourier transform. The frequency components of a sound can be identified. Sound when "identified" is matched to the database. The tools used are ASP.NET platform, C# language and SQL.

Reference [7] shows how to use the speech recognition in the windows vista. Microsoft uses the concept of speech technology, synthesizers and recognizers. In short speech synthesizers is the technology that responds to user's voice. JAWS is another speech synthesizer tool widely equipped to read information on the screen and capable of converting text messages on the screen to audio output formats. The latest version of Opera browser can be controlled by voice commands and will read pages aloud. If we make use of Java API, two core speech technologies are supported through the Java Speech API: speech recognition and speech synthesis which require the application such as Live Connect, Speech Synthesizer, Chrome and XPCOM component.

Reference [8] discusses HMM process where the speech samples containing unwanted signals and background noise are removed by end point detection method. The steps of Mel frequency Cepstral Coefficients (MFCCs) calculation are performed– framing, windowing, Discrete Fourier Transform (DFT), Mel frequency filtering, logarithmic function and Discrete Cosine Transform. Firstly, five audio files are recorded with the help of computer. Each audio file contains ten different pronunciation audio files. So, there are total of fifty audio files are recorded in speech database. After pre-processing stage is finished, the speech samples are extracted to features or coefficients by the use of Mel Frequency Cepstral Coefficient (MFCC). Finally, these MFCC coefficients are used as the input of Hidden Markov Model (HMM) recognizer to classify the desired spoken word.

Reference [9] shows a system that takes speech as input at run time through a microphone and processed the sampled speech (using PMC) to recognize the uttered text. The recognized text can be stored in a file. The system was developed on android platform using eclipse workbench. Process involved the conversion of acoustic speech into a set of words performed by software component. Speech recognition system has been divided into several blocks: feature extraction, acoustic models database built based on the training data, dictionary, language model and the speech recognition algorithm. Analog speech signal was first sampled on time and amplitude axes, or digitized. Samples of speech signal were analysed in even intervals. Speech recognition used a technique based on hidden Markov models (HMM).

Reference [10] discusses the three types of approaches of speech conversion: Acoustic phonetic approach, Pattern Recognition and Artificial intelligence approach. Acoustic phonetic approach uses knowledge of phonetics & linguistics to guide search process, usually some rules which are defined expressing everything or anything that might help to determine which sequences are permitted. Pattern Recognition Approach has two steps: 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. Artificial Intelligence Recognition

Approach is a combination of the acoustic phonetic approach and the pattern recognition approach. In the AI, Neural networks implemented a system used to classify sounds. The basic idea is to compile and incorporate knowledge.

Reference [11] uses vectors for speech representation and comparison. Feature extraction is the first step. The main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. Mel Frequency Cepstral Coefficients (MFCC) technique was used to extract features. It includes basic steps: frame blocking, windowing, FFT, Mel Filter Bank Processing, Mel frequency wrapping and cepstrum. Speech pattern representation could be in the form of a speech template or a statistical model (e.g. HMM).

Reference [12] describes a system equipped a speech-to-text module using isolated word recognition with a vocabulary of ten words (digits 0 to 9) and statistical modelling (HMM) for machine speech recognition. In the training phase, the uttered digits are recorded using 16-bit pulse code modulation (PCM) and saved as a wave file using sound recorder software.
The MATLAB software's wavered command has been used to convert the .wav files to speech samples. From the LPC coefficients, the weighted cepstral coefficients and cepstral time derivatives are obtained, which form the characteristic vector for a frame. Then, the system performs vector quantization which forms the observation sequence. For each word in the vocabulary, the system builds an HMM model and trains the model during the training phase. The training steps, from VAD to HMM model building, are performed using PC-based C programs. The resulting HMM models are loaded onto an FPGA for the recognition phase.
In the recognition phase, the speech is acquired vigorously from the microphone through a codec and is stored in the FPGA's memory. The uttered word is recognized based on maximum likelihood estimation.

Reference [13] implements a speech-to-text system using isolated word recognition with a vocabulary of limited words (recognized with SDK 5.1) and statistical modelling (HMM) for machine speech recognition. In the training phase, the uttered digits are recorded using 16-bit pulse code modulation (PCM) and saved as a wave file using sound recorder software. We use the MATLAB software's wave read command to convert the .wav files to speech samples. The speech is separated from the pauses using voice activity detection (VAD) techniques. The system performs speech analysis using the linear predictive coding (LPC) method. From the LPC coefficients we get the weighted cepstral coefficients and cepstral time derivatives, which form the feature vector for a frame. Then, the system performs vector quantization which forms the observation sequence. We implemented these steps with C# programming, which was executed based on the algorithms. It will supply the Hidden Markov Model (HMM). It will also supply the dedicated tools for Quartus II software.

Reference [14] describes a process to extract features by using Mel Frequency Cepstral Coefficients (MFCC) from the speech signals of isolated spoken words and Hidden Markov Model (HMM) method is applied to get the recognized spoken word. The speech database is created by using MATLAB. Then, the original speech signals are pre-processed and these speech samples are extracted to the feature vectors which are used as the observation sequences of the Hidden Markov Model (HMM) recognizer. The steps of MFCC calculation are– framing (signal divided into frames), windowing (to reduce discontinuities at the start and the end of the frame), Discrete Fourier Transform (used as the Fast Fourier Transform algorithm converting each frame of N samples from the time domain into the frequency domain), Mel frequency filtering (corresponds to better resolution at low frequencies and less at high), logarithmic function (gives an absolute magnitude operation and discards the phase information, making feature extraction less sensitive to speaker dependent variations) and Discrete Cosine Transform (converts the Mel-filtered spectrum back into the time domain since the MFCC are used as the time index in recognition state). Finally, these MFCC coefficients are used as the input of Hidden Markov Model (HMM) recognizer to classify the desired spoken word (HMM model is used for speech processing).

Reference [15] discusses about the various decoding methods of speech which are acoustic phonetic method, pattern recognition method, and artificial intelligence approach. Pattern recognition mainly incorporates two steps, including pattern comparison and pattern training. Two approaches for pattern recognition are stochastic approach and template approach. The other decoding method used is Acoustic Phonetic. This method is based on the process of locating sounds and speeches (foundation of the acoustic phonetic approach). The third approach for decoding is Artificial Intelligence. It can be understood as the combination of the pattern recognition approach and acoustic phonetic approach. It includes designing of recognition algorithm, demonstration of speech units, and representation of appropriate inputs. Among all methods of speech recognition, artificial intelligence is the most efficient method.

Reference [16] propose a system which uses the data-driven approach to phoneme classification, thus attempting to solve the problem of inter and intra speaker speech variability, by the use of a large speech database. It also has the ability to generate decision trees using any combination of features (parametric or acoustic-phonetic). The four feature extraction modules are: MFCC, DCT, FEATURE and TRAJ. Modules: MFCC and feature extract features in the time domain while modules DCT and TRAJ extract features (from the auditory model) in the frequency domain. The recognition is performed at the frame level and the performance is evaluated by comparing each classified frame against the reference frame derived from the hand labelled data. This procedure allows the correct identification of substitutions and insertions per frame.

III. Types of Approaches

A. *Template Based Approach*

Template-Based Approach has been used for speech-to-text conversion. The output from the signal processing module of a speech recognizer is a sequence of feature vectors. One first builds a collection of reference templates, each itself a sequence of feature vectors that represents a unit (usually a whole word) of speech to be recognized. Then, the feature vector corresponding to the current utterance is compared via some distance measure. The template-based approach has produced favourable results for small-dictionary applications, mainly for IWR.

B. *Pattern Recognition Approach*

Pattern Recognition Approach has two steps: training of speech patterns and recognition. In the parameter measurement phase, a sequence of measurements is made on the input signal to define the “test pattern”. The unknown pattern is compared with sound reference provided to the database.

C. *Knowledge-Based Approach*

In Knowledge based systems, the linguist attempts to describe and quantify the acoustic events, in the form of production rules into phonetic description. The knowledge itself can then be induced from examples in the agreed structure. Thus the acoustic-phonetic rules are moved to the machine memory. Recognition results on three broad phonetic classes, namely semi-vowels, for combination of feature sets, for speaker dependent- independent recognition.

D. *Dynamic Time Warping-Based Approach*

Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in speed. For instance, similarities in walking could be detected using DTW, even if one person was walking faster than the other. DTW has been applied to temporal sequences of video, audio and graphics data — indeed, any data which can be turned into a linear sequence can be analysed with DTW. A well-known application has been automatic Speech Recognition, to cope with different speaking speeds. Other applications include Speaker Recognition and Signature Recognition.

IV. Application Models

A. *Hidden Markov Model (HMM)*

Modern general-purpose speech recognition systems are based on Hidden Markov Models, an application of Template and Pattern Based Approaches. 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. A piece wise signal is similar to a Template. In a short time-scale (e.g., 10 milliseconds), speech can be approximated as a stationary signal. Speech can be thought of as a Markov Model for many stochastic purposes.

B. *JAWS*

JAWS (Job Access With Speech) is a computer screen reader program for Microsoft Windows that allows blind and visually impaired users to read the screen either with a text-to-speech output or by a refreshable Braille display

C. *Sound Recording Software (MATLAB)*

This is an audio recording software developed using MATLAB. We could see the usage of HMM in this application. This model tries to implement speech recognition along with HMM using waveforms.

V. TABLE I

A Comparative Table

Approach/Model	***Voice/Speech***	***Strengths***	***Weakness***
Template-Based Approach	Each word is stored as a vector.	Uses accurate word models and segmentation. Errors due to segmentation or classification into	Separate template for each word brings dependency over smaller units like phonemes. Non-linear time alignment is crucial. Ex: inevitable different speaking rates. Even for a same person, we

		smaller variable units such as phonemes can be avoided.	have different speaking rates for the same word uttered. Reliability determines the boundary of the word.
Pattern Recognition Approach	Speech is measured to define a "TEST PATTERN"	No speech-specific knowledge is used in the system; hence, this method is relatively insensitive to the choice of vocabulary of words, syntaxes, and semantics. Because the system is insensitive, it can be used to recognize an entire phrase, whole word, sub-words.	The reference patterns are sensitive to the speaking environment and transmission characteristics of the medium used to create the speech; this is because speech is affected by transmission and background noise. The computational load for pattern classification is generally linearly proportional to the number of patterns being trained or recognized; hence, computation of a large number of phrases becomes prohibitive or difficult.
Knowledge-Based Approach	Speech is hand-coded into a system.	This model is explicitly modelling variation in speech. The speech recognized is stored for future comparisons and results.	Space Consuming. Heavy memory usage.
Dynamic time warping- Based Approach	Speech is measured to get frequencies of words.	It is Language Independent. Easy to train the templates for future recognitions. Works well for smaller templates.	It is not suited for continuous speech recognition patterns. It requires the computation of the words starting and ending points so that frequency is easily noted.
Hidden Markov Model	Speech is considered to be piecewise stationary signal or a short-time stationary signal.	Good abstractions for sequences compared or recognized. Easy to decode the sequences recognized.	It is not completely automatic. It needs a manual mark up.
Jaws	Only output	It allows a blind person to create a document using a word processor like MS Word. A blind person can now read any article on the internet and also write email.	Because visually impaired people only listen to a screen reader reading the text displayed on the screen, they don't usually have the chance to know the correct spelling of a certain word especially when it's not that common like medical terms etc. Screen readers use a computerized voice. Some companies are doing their best to create speech synthesizers that can mimic how human read a sentence.
MATLAB Sound Recording Software	Speech is divided into frames.	Fast Speed Easy Extension	Audio recorder is not intended for long, high-sample-rate recording. Audio recorder uses system memory for storage and does not use disk buffering or storage. When you attempt a large recording, your MATLAB performance sometimes degrades over time.

VI. CONCLUSIONS

The analysis of different approaches shows us how each of the process is carried out and how better methods can be implemented. The comparisons between the various approaches and the models that are derived or used to implement the approaches stated , gives us an idea of which process to be used based on the input. Each process has its own pros and cons. Considering the various environmental aspects, surroundings, application of the method, we can easily map which one to use. Considering all the references, we could say that using of HMM model in the modern day applications is the best feasible method.

Hence, the comparative table could be used to get a better approach processed for any given type of input.

ACKNOWLEDGMENT

We would like to thank TEQIP (Technical Education Quality Improvement Programme) Phase-II for giving us an opportunity to carry out our survey.

REFERENCES

[1] Susanne Wagner (Halle), "*Intra lingual speech-to-text-conversion in real-time: Challenges and Opportunities*", EU-High-Level Scientific Conference Series, 2002

[2] Adam L. Buchsbaum and Raffaele Giancarlo, "*Algorithmic Aspects in Speech Recognition: An Introduction*", Association for Computing Machinery, Inc., 1515 Broadway, New York – 2005

[3] Kumbharana, Chandresh K., "*Speech Pattern Recognition for Speech to Text Conversion Thesis*", Saurashtra University

[4] Philippe Drew, David Rybach, Thomas Deselaers, Morteza Zahedi, "*Speech Recognition Techniques for a Language Recognition System*", RWTH Aachen University, Germany – 2008

[5] Nitin Washani and Sandeep Sharma, "*Speech Recognition System*", International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 18, April 2009

[6] *Voice based automatic enquiry system* 2010

[7] Sadaoki Furui, "*Speech-to-Text Conversion*", IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 4, July 2011

[8] Jan Novotny, Pavel Sovka, Jan Uhlir, "*Speech-to-Text conversion and reducing noise*", Radio-engineering, Vol. 13, No. 1, April 2011

[9] B. Raghavendhar Reddy, "*Speech to Text Conversion using Android Platform*", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, www.ijera.com Vol. 3, Issue 1, January -February 2013

[10] Preeti Saini, "*Automatic Speech Recognition: A review*", International Journal of Engineering Trends and Technology- Volume 4 Issue2- 2013

[11] Alaa Hassan Mahmoud, Salma Alzaki Ali, "*Speech to text conversion*", A thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Bachelor in Software Engineering – 2014

[12] Miss. Prachi Khilari, "*A Review on Speech to Text Conversion Methods*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 7, July 2015

[13] Rajesh Makhijani, "*Speech recognition system*", International Journal of Engineering Sciences & Emerging Technologies, ISSN: 2231 – 6604 Volume 6, Issue 3, Dec. 2015

[14] Su Myat Mon, Hla Myo Tun, "*Speech-To-Text Conversion (STT) System Using HMM*", International Journal of Scientific & Technology Research Vol. 4, Issue 06, June 2015

[15] Khaled M. Alhawiti, "*Advances in Artificial Intelligence*", and World Academy of Science, Engineering and Technology, International Journal of Computer, Vol: 9, No: 6, 2015

[16] A.Samouelian, "*Knowledge Based Approach to Speech Recognition*", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2015