Systematic Review On Speech Recognition Tools And Techniques Needed For Speech Application Development

Lydia K. Ajayi, Ambrose A. Azeta, Isaac. A. Odun-Ayo, Felix.C. Chidozie, Aeeigbe. E. Azeta

Abstract: Speech has been widely known as the primary mode of communication among individuals and computers. The existence of technology brought about human computer interface to allow human computer interaction. Existence of speech recognition research has been ongoing for the past 60 years which has been embraced as an alternative access method for individuals with disabilities, learning shortcomings, Navigation problems etc. and at the same time allow computer to identify human languages by converting these languages into strings of words or commands. The ongoing problem in speech recognition especially for under-resourced languages has been identified to be background noise, speed, recognition accuracy, etc. In this paper, in order to achieve a higher level of speed, recognition accuracy (word or sentence accuracy), manage background noise, etc. we have to be able to identify and apply the right speech recognition algorithms/techniques/parameters/tools needed ultimately to develop any speech recognition system. The primary objective of this paper is to make summarization and comparison of some of the well-known methods/algorithms/techniques/parameters identifying their steps, strengths, weaknesses, and application areas in various stages of speech recognition systems as well as for the effectiveness of future services and applications.

Index Terms: Speech recognition, Speech recognition tools, Speech recognition techniques, Isolated speech recognition, Continuous speech recognition, Under-resourced languages, High resourced languages.

1 INTRODUCTION

Over the years, a lot of research have been done on developing vocally interactive computers for voice to speech realization which has been of great benefits and this speech synthesis can be classified as either Automatic Speech Recognition also called Speech to Text Synthesis and Text to Speech Synthesis [1]. The aim of developing any speech recognition system is to allow computer systems to accurately recognize human speech and the best way to accomplish this aim is to be able to have deeper understanding on the right approaches, tools and techniques that can be applied in developing these recognition systems. Automatic speech recognition is also known as Speech to Text that does the conversion of speech signal into a textual information i.e. a sequence of spoken words using an algorithm that is implemented by a software or hardware module into a text data[1] [2]. Automatic speech recognition by machine has been a research goal for decades. Even in the science world, human mimics has always been understood by computers. The main idea for developing speech recognition system was generated because it brings about conveniences as humans can easily interact with a computer, machine or robot through the aid of speech/vocalization rather than using difficult instructions [3].

- *Lydia K. Ajayi is currently pursuing PhD degree program in Computer Science in Covenant University, Nigeria. E-mail: lydia4reel@gmail.com*
- *Ambrose A, Azeta is currently a Professor of Computer Science in Covenant University, Nigeria.. E-mail: ambrose.azeta@covenantuniversity.edu.ng*
- *Isaac A. Odunayo is currently a Doctor of Computer Science in Covenant University, Nigeria.. E-mail: isaac.odunayo@covenantuniversity.edu.ng*
- *Felix C. Chidozie is currently a Doctor of Political Science in Covenant University, Nigeria.. E-mail: felix.chidozie@covenantuniversity.edu.ng*
- *Aeeigbe E. Azeta works at PTTIM in FIIRO, Nigeria.. E-mail: felix.chidozie@covenantuniversity.edu.ng*

ASR is an application that consistently explores computational capability enhancement advancement [4]. Its main task is to decode the most probable word sequence either isolated words or continuous speech based on audio signal with speech [5] ASR system is about optimization by improving accuracy that have to do with noisy and reverberant environments, improving throughput by allowing batch processing of speech recognition task to be executed efficiently for multimedia search and retrieval and then improving latency which also include achieving real time performance [4].

1.1 Classification of Speech Recognition

Speech Recognition is classified based on three (3) different modes which are:

1.1.1 Classification Based on Utterances

Isolated Word Recognition. The recognition of isolated words usually require every utterances to be quiet on both side of sample windows. It accepts single words/ utterances at a time with having both “Listen and Non Listen state”. Isolated word recognition is the process in which word is surrounded by some sort of pause [6] [7].

Continuous Word Recognition. The recognition of continuous speech allows user of the system to talk freely and naturally while the computer determine the content(s). Its process involves allowing words to run into each other and have to be segmented [1] [6] [7] does not take pauses between spoken words. There are difficulties in creating continuous speech recognizer as they utilize special method and unique sound in order to determine the speaker utterance boundaries.

Spontaneous Word Recognition. The recognition of spontaneous word at the basic level can be envisaged a natural sounding and not rehearsed. It is known as a human speech to speech system with spontaneous speech characteristics which has the capability to handle different

words and various natural speech feature such as running together of words.

Connected Word Recognition. Recognition of connected words is also similar to isolated words but the difference is it allows the running together of separate utterances which is word vocalization that represent a single meaning to a computer with minimal pauses between them [8] [9].

1.1.2 Classification Based on Speaker Model

Every speaker has a unique voice due to unique physical body and personality and it is broken down into "3" different categories.

Speaker Dependent Model. These systems are systems that are developed for a particular type of speaker and they have a generally more accurate level for that particular speaker [9] [10]. It can be less accurate when another speaker tries to use the system but the advantage is that they are relatively cheaper, accurate and easier to develop but the drawback is that they aren't flexible compared to speaker independent systems.

Speaker Independent Model. The recognition of continuous speech allows user of the system to talk freely and naturally while the computer determine the content(s). Its process involves allowing words to run into each other and have to be segmented [1] [6] [7] does not take pauses between spoken words. There are difficulties in creating continuous speech recognizer as they utilize special method and unique sound in order to determine the speaker utterance boundaries.

Speaker Adaptive Model. The recognition of spontaneous word at the basic level can be envisaged a natural sounding and not rehearsed. It is known as a human speech to speech system with spontaneous speech characteristics which has the capability to handle different words and various natural speech feature such as running together of words.

1.1.3 Classification Based on Vocabulary

The size of speech recognition system vocabulary can majorly affects the ASR system recognition based on complexity, processing and rate. Classification of this vocabulary is categorized into;

Small Vocabulary. This is 1-100 words or sentences systems.

Medium Vocabulary. This is 101-1000 words or sentences.

Large Vocabulary. This is 101-1000 words or sentences.

Very Large Vocabulary. This is greater than 10,000 words or sentences.

In order to develop any of this mode of speech recognition for higher accuracy, speed and latency etc. starting from classification based on utterances (Isolated word recognition, continuous word recognition, etc.) to classification based on speaker model (speaker dependent model, speaker independent model. etc.) to classification based on vocabulary (small vocabulary, medium vocabulary, etc.), There is a need to understand the functioning of speech recognition system especially for under-resourced languages which are known as languages that lacks orthography or unique writing system, limited web presence like, transcribed speech data, monolingual corpora, bilingual electronic dictionaries, pronunciation dictionaries, and vocabulary list. They are also called low-density, low-data, low-resourced or resource poor languages [1] with the appropriate techniques/algorithms needed for each steps. Most existing approaches produce models that are either generalized, inappropriate or inefficient for developing speech recognition systems. These speech recognition systems exhibits low recognition accuracy, speed, latency, etc. This brought about the existence of speech recognition tools, techniques and algorithms. Most tools/techniques/algorithms have been employed having unknown drawbacks which later generates setback for the speech systems. In order to be able to achieve high recognition accuracy, speed, latency, etc., individual involved in the aspect of research needs have a deep insight on the steps, drawbacks, strengths, etc. of each tools, techniques and algorithms to be used in order to choose optimal best needed for their project work.

2 LITERATURE REVIEW

Unsupervised adaptation methods was used to develop an isolated word recognizer for Tamil language which is an under-resourced language which makes unsupervised approaches to be attractive in this context. Similar and extended approaches was also introduced for Polish language [11] and Vietnamese Language [12] but exhibited low accuracy. [13] developed an HMM based Tamil Speech Recognition which is based on limited Tamil words but also exhibited low recognition accuracy. Automatic speech recognition for Tunisian dialect, an under-resourced language developed by [14] where he compared HMM-GMM model using MFCC, and HMM-GMM with LDA. These approaches gave a WER of 48.8%, 48.7%. [15] developed a standard Yoruba Isolated ASR system using HTK which has the capability to act as an isolated word recognizer which is spoken by users based on previously stored data. The system adopted syllable-based approach using 6 native speakers speaking 25 bi-syllabic and 25 tri-syllabic words under acoustically-controlled room based on HMM and MFCC. The overall accuracy recognition ratio gave 76% and 84% for bi-syllabic and tri-syllabic words. [12] also developed a Vietnamese ASR system which is an under-resourced language and achieved a syllabic error rate of 16.8% and 16.1% on evaluation set. A continuous speech recognition for Indian languages (Tamil, Telegu and Marathi) was also developed by IIIT Hyderabad using HMM technique and Sphinx 2speech toolkit. The system attained a word error rate suing three different error rate (WER), of 23.2%, 20.2% and 28% for the three languages respectively. From the review given above, it can be deduced that recognition accuracy, speed, noise, latency etc., is still a problem in speech recognition especially for under-resourced languages and to combat or avoid these problems, there is a need to understand the right techniques based on the mode of application classification to be developed.

3 STATEMENT OF PROBLEM

The challenges facing the development of speech recognition system of today amounts to recognition accuracy level, throughput level, latency level, etc. which affects the overall quality of the speech system. For example every speaker has their unique way of pronouncing words through their accents,

styles and speech rates which can affects the quality of a speech recognition systems. Another popular challenging is noise. Evaluation percentage gotten for accuracy level for most speech recognition system are not near perfect. To counter these aforementioned problems, all speech recognition technologies (tools, techniques, algorithms, or approaches) needs to be reviewed and studied so as to have a deep insight into their steps, application areas, strengths, drawbacks, etc. This aids in selecting the optimum best to be applied for development of speech recognition system for practical application. Research work done in speech recognition only gave a broad survey on the latest trend on speech recognition but didn't properly align the steps, limitations, strengths, how they work etc. of each tools/algorithms/approaches used and how and when it is proper to be applied to future work so as to achieve a good quality speech system especially in the area of under-resourced languages. There is therefore need to introduce a qualitative/comparative analysis of different algorithms/tools/techniques.

Therefore, the objectives of this paper are as follows:

(1) To review the concepts of speech recognition and recapitulate/matchup different techniques with focus on under-resourced languages.
(2) To align the general issues facing speech recognition in relation to developing a speech recognition system.
(3) To make a comparative analysis of speech recognition techniques, approaches and tools detailing their steps, strengths, limitations/weaknesses, application areas, etc.

4 CURRENT RESEARCH ISSUES IN SPEECH RECOGNITION TO DEVELOPING SPEECH RECOGNITION SYSTEM

The development of an SRS comes with a lot of challenges which are detailed below:

4.1 Speech Variability

Speech variability is a challenging issue in speech recognition system which is divided into inter speaker variability and intra speaker variability which all incudes different accents, contexts, voices, styles, and speech rates.

4.2 Recognition Units

This is also another challenging issue which includes words and phrases, syllables, phonemes, diaphones and triphones in speech.

4.3 Language Complexity

Language complexity includes vocabulary size and difficulty which also affects developing speech recognition systems.

4.4 Ambiguity

This includes word boundaries, homophones, syntactic and semantic ambiguity.

4.5 Environmental Conditions

Environmental conditions include background noise or several people speaking simultaneously which may be coherent or not. Variations due to background noise increases error rates in speech recognition systems. Noise leads to low recognition accuracy as they can be reverberant or serious.

4.6 Low Throughput

Low throughput is also another research issue which involves system efficiency which is needed in today's energy limited and form factor limited technology. Improving throughput will allow the batch processing of speech recognition tasks to execute efficiently well [4].

4.7 Low Latency

For a speech recognition system to be ultimately useful especially in multiple speaker environment/applications, speech diarization and speech recognition is needed. Speech diarization involves the ability for a system to know who is speaking which is reasonable for human consumption. Most current research systems doesn't consider speaker diarization and segmentation which actually helps in improving latency. In order to achieve high rate of recognition accuracy, speed, latency, robustness throughput etc. we have to have a deeper understanding on the functioning of speech recognition identifying each techniques with their strength, weaknesses/limitation, areas where the techniques can be applied, together with the speech recognition tools.

6 FUNCTIONING OF A SPEECH RECOGNITION SYSTEM WITH ITS TECHNIQUES

Functioning of speech recognition systems constitute different phases which are aligned in fig (1) below but the focus of this research is on the processing/feature extraction and acoustic modeling detailing their techniques/algorithms. The ASR system components includes the feature extraction component, the inference engine, acoustic modeling, language modeling, pronunciation modeling, and decoder [4]. Speech recognition process starts with the creation of an utterance that consists of sound wave. These sound waves are then captured by microphone which then converts it into electrical signals. The electrical signals are then converted into digital form for speech recognition and then converts the speech signal into discrete sequence of feature vectors with an assumption that it contains only relevant information about the given utterance which is needed for correct recognition. The feature extraction helps to extract irrelevant information e.g. noise for correct classification. The recognition component finds the optimum best match in the knowledge base for incoming feature vectors [16]. Figure 1 gives the architecture of an ASR system.

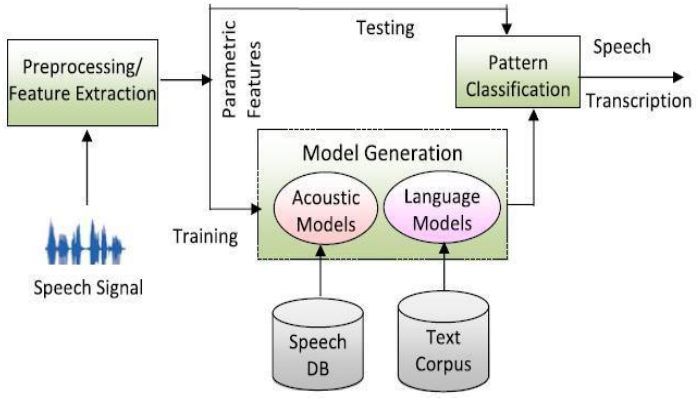

***Fig 1**. A schematic architecture of a speech recognizer decoding a single sentence (Source: [9])*

5.1 Preprocessing/Feature Extraction

This is the first step in Speech recognition where the recorded signal is an analog signal. Analog signal cannot be transferred and easily recognized by the ASR systems. Therefore the analog speech signals needs to be transformed into a digital signal that the ASR system can easily recognized. The digital signals are then moved to the first order filters in order to spectrally flatten the signal which performs energy signals increment at higher frequency. This step is called the preprocessing stage. Also, this step is one of the most important part of speech recognition and helps to separate one speech from another due to the fact that every speech has unique individual characteristics embedded in their utterances [17]. It extracts features by identifying the component of an audio signal and discarding the irrelevant information like noise, emotion, etc. i.e. keeps relevant speech information and discards irrelevant ones. These utterances or features can be extracted using a wide range of feature extraction techniques but the commonly used techniques are;

5.1.1 Mel-frequency Cepstrum Coefficients (MFCC)

MFCC are designed using human auditory system. This method aids in finding our features but the disadvantage it has is that the values are not very robust in the presence of additive noises which introduces normalization values in order to reduce the influence of noise [18].

5.1.2 Linear Predictive Coding (LPC)

In LPC, the basic idea behind this technique is the approximation of speech sample which can be as linear combination of past samples which provides observation vectors of speech.

5.1.3 Principal Linear Prediction (PLP)

The Principal Linear Prediction (PLP) is used in the description of the psychophysics of human hearing system in a more accurate way. It is very similar to LPC which is based on short-spectrum of speech but modifies the speech based on multiple psychophysically based transformations [19].

5.1.4 Linear Discriminate Analysis (LDA)

It is a well-known algorithm for improving discrimination and compression of information contents. It is a supervised linear map that depends on Eigen vector, with its nonlinear feature extraction method.

5.1.5 Relative Spectral Processing (RASTA)

Rasta involves linear filtering of the trajectory short power spectrum of noisy speech signal. It is basically used for noisy speech by finding feature in noisy data through filtering.

5.1.6 Discrete Wavelet Transform (DWT)

DWT allows permission of high frequency events identification with improved temporal resolution which is centered on signal analysis using varying scales in frequency and time domains. Table (1) gives a detailed comparison of different feature extraction techniques with their steps, benefits and drawbacks with properties for implementation which is needed for selecting the optimum best feature extraction techniques based on task at hand.

Table 1

A Comparative Analysis of Feature Extraction Techniques/Algorithms.

S/N	Algorithms/Techniques	Computational Steps	Computational speed	Property needed for implementation	Strengths	Drawbacks
1	MFCC	Pre-emphasis (framing and blocking), windowing, Discrete Fourier transform DFT, mel filter bank and log, Discrete Cosine Transform (CCT).	High	Implementation of Fourier Analysis is computed by Power Spectrums.	.It can perform better both in isolated and continuous speech recognition tasks for finding features. Low frequency region can efficiently be denoted better compared to high frequency region	In the presence of additive noises, MFCC values are not very robust which makes it require normalization. It might not be well- suitable for generalization.
2	LPC	Inputting Speech signal, frame blocking, windowing, autocorrelation analysis, analysis of LP based on Levinson-Durblin which outputs the LP feature vectors.	High	It is a static feature extraction technique and useful for feature extraction at lower order due to its 10-16 lower sequence coefficient.	It provides a spectral analysis with a fixed resolution along a subjective frequency scale (Mel frequency scale). It is used for speech reconstruction and helps to efficiently extract vocal tract properties.	The frequencies are weighted equally on a linear scale in which human ear frequency sensitivity is almost logarithmic. It suffers from aliased autocorrelation coefficients. Not well suited for generalization.
3	PLP	Input windowed speech, FFT, frequency warping, equal loudness pre-	High	Combination of critical bands, compression, equal loudness, pre-	Identification achieved by this algorithm is better than LPC as it suppresses speaker dependent	Sensitivity to any changes in formant frequency.

		emphasis, compression of amplitude, DCT, Cepstral recursion, Dublin recursion and gives the output PLP feature vectors.		emphasis and loudness intensity for speech feature extraction	information. Noise robustness, channel variations and microphones. The algorithm accurately constructs auto-regressive noise.	
4	LDA	Computation within class and between class scatter matrices, computation of Eigen vectors and corresponding Eigen values for the scatter matrices, and then outputs the LDA components	Medium	It is a non-linear algorithm, fast, supervised linear map and Eigen-vector based.	Performance is better compared to PCA for classification. It is useful in handling tasks where the class frequencies are unequal and also performs better on randomly generated test data.	The performance ratio is less in distribution that is non-Gaussian as it will not be able to preserve any complex data structure which is ultimately needed for classification.
5	RASTA	Compression of the spectral values of input speech signal by non-linear compression rule, filtering operation and expansion after filtering.	Medium	Adequate for noisy speech.	Beneficial for noisy speech data and reverberant data feature extraction.	It increases data dependency based on its previous context.
6	DWT	Input speech signal, low-pass filter or high-pass filter and outputs DWT parameter	High	Finite set of observations analysis over set of scales	Ability to concurrently mine information from transient signals in both frequency and time domains.	It does not fulfill the direct use of in parameterization. It does not make provision for adequate number of frequency bands for efficient speech signals.

Notes: In Table 1, different algorithms were compared based on computational steps, computational speed, implementation property, strengths and drawbacks. From the table, it can be deduced that MFCC works better for noisy speech data compared to other techniques.

5.2 Acoustic Models

Acoustic Models is the main component of an ASR system which is developed for detecting spoken phoneme [3]. This stage is also a fundamental part of ASR system where acoustic information and phonetics connection are established. This component plays a critical and important role in system performance as it is also responsible for computational load. The acoustic modeling component is divided into speaker recognition and identification [18]. The speaker recognition extract the speaker characteristics in the acoustic signal and the speaker identification part automatically identifies who is speaking based on individual information integrated in the speech signal by comparing the speech signal from an unknown speaker which has been trained with a number of speakers. The training helps to establish the co-relation between basic speech units and the acoustic observations. The training of the system requires the creation of a pattern representative for class feature with the use of one or more patterns that corresponds to speech sounds of the same class feature [9]. Many models are now available that can be utilized in speech recognition process;

5.2.1 The Acoustic-Phonetic Approach
This approach has been researched for over 40 years which is finding speech signals and the appropriate labels to these

sounds. It is also known as a rule-based approach that uses linguistics knowledge and phonetics in order to guide search process.

5.2.2 Template-Based Approach
This approach makes use of a set of pre-recorded words known as template against the unknown speech in order to find the optimum match. The underlying idea is simple as it is the collection of prototypical speech patterns that are stored as reference patterns which represents dictionary of candidate words.

5.2.3 Hidden Markov Model (HMM)
It is grouped under the statistical method and is a widely used and popular statistical approach. HMMs are simple networks where system to be modeled based on the assumption that they are markov process with unknown parameters. This model helps to generate speech from unknown utterances and then make the comparison that the unknown utterances was generated by each mode. HMM are widely used and recognized due to the fact that it is simple, can be trained automatically and computationally feasible.

5.2.4 Gaussian Mixture Model (GMM)
This model is a multivariate Gaussian model which is a weighted sum of Gaussian distributions that is used to assign a likelihood score to an acoustic feature vector observation. It is a probabilistic model that takes almost all information which is produced from a blend of a limited number of obscure parameters. It has been known to perform best in languages with low data training sets compared to other models. They are used as probability distribution features such as vocal-tract spectral features in a speech recognition or emotion recognition.

5.2.5 Expectation Maximization (EM) Algorithm
The EM approach can be used to get maximum probability. When data is incomplete/unexpected, it evaluates for structure elements and the method/process is repeated many times until the optimum maximum task is achieved.

5.2.6 Dynamic Time Warping (DWT) Algorithm
This is an approach that has been used for a long time in speech recognition and allows non-linear mapping of one signal to another by minimizing the distance the two signals [19] and also exhibits non-linear time normalization effect. It basically finds similarities between two different signals which vary between time and speech. Any type of machine matching is done with certain restriction.

5.2.7 Vector Quantization (VQ)
This approach is a classical quantization technique that allows modeling of PDFs by distribution of prototype vectors. It functions by dividing a large set of points into groups having approximately same number of points that is closest to them.

5.2.8 Support Vector Machine (SVM)
SVM is also an acoustic approach that is mainly used for speaker verification and identification. It is a twofold classifier that models choice limit between two classes as an isolated hyper-plane. It is either a knowledge-based Approach which is also called the rule-based approach as it utilize information regarding either linguistic, phonetic and spectrogram. Expert knowledge about speech variation is hand-coded into the system which then extract features from the speech and performs system training to generate a set of production rules automatically from the samples [18].

5.2.9 The Artificial Based Approach
A neural network is also a reasoning model that is based on human brain as the human brain consists of neurons. The human brain consists of billions of neurons and 60 trillion connections called synapses between them. Using many neurons together will allow the brain to work effectively and performs faster than modern computers. The artificial neural network consists of neurons that are analogous to the human brain neurons. The ANN neurons are interconnected together by weighted links. ANN learning can either be supervised, unsupervised or reinforced. The supervised learning involves when the data you feed your algorithm is tagged to help make logical decisions like face recognition, perceptron. The Unsupervised learning is to discover similar groups that share common properties within data using clustering. In reinforced learning, data sets are not usually given but the data are generated by agent's interaction with the environment [20].
The Comparative analysis of acoustic model algorithm is shown in Table (2).

Table 2
A Comparative Analysis of Acoustic Models/Approaches.

S/N	Acoustic Approaches	Computational Steps	Strengths	Drawbacks	Application Areas
1	Acoustic-Phonetic Approach	Analysis of speech spectral with the combination of feature detection, segmentation and labeling phase and determination of valid words or string of words from labelling and segmentation phase.	Lay out possibility and rules application for determination of sequence permission. The rule governing phonetic variability is straightforward and easily read by the machine.	Poor performance of difficulty in expressing rules to improve the system which makes it not widely utilized for commercial applications.	Isolated Digit Recognition. Speaker Independent.

2	Template based Approach	prototypical speech patterns collection which are stored as reference patterns which represent dictionary of candidate words, carrying out of speech recognition by matching unknown utterance with each reference templates and selection of optimum matching pattern	It has an advantage of using perfectly accurate word models. It exhibits simplicity and better for discrete words. Error occur due to segmentation/ classification of phonemes can be avoided.	Continuous Speech recognition is impossible	Small Dictionaries. Isolated Word Recognition Speaker dependent.
3	HMM	Collection of states connected by transitions, a transition is taken into a new state with generation of one output symbol in that state.	The vocabulary size of this model is very large. It has an accurate mathematical framework. It is widely used and easily accessible, easily be implemented, flexible and use unsupervised learning method. It adequately does the temporal and spectral variations analysis of speech signals, and can recognize and decode continuous speech input efficiently.	It can't be applied on low resourced language (low training data) It requires large training data sets to perform accurate speech recognition.	Isolated, Continuous Speech Recognition. Speaker independent High resourced language.
4	GMM	Consideration of component covariance structure, specifying initial conditions and implementing regularization	Easy to handle for computation. More flexible, accurate, easy to fit into a given multivariate data set. Performs better in smaller data sets. Recognizes Spectral features compared to HMM Requires less preparation and test data.	Direct maximization of trained data is impossible. It can incorrectly the more common choice as it uses probability.	Isolated, Continuous Speech Recognition. Speaker independent Low resourced language.
5	EM	Structure elements evaluation when data is incomplete or unexpected, repetition of method is done many times in order to find the maximum probability	It is powerful and robust which is used in iteratively finding homologous data set and choosing the optimum best.	It is critically dependent on any initialization	Isolated, Continuous Speech Recognition. Speaker independent Low resourced language
6	DTW	Matching of index from first sequence with one or more sequences and vice versa to get the DTW output	Continuity is less important as even with missing information as it can match sequence. There is reliable time alignment of segments of two signals which is reference and text pattern.	Limited number of templates. It requires plenty time for complex computational task. It requires unknown variant speech signal for training.	Connected Speech Recognition. Isolated word recognition Character Recognition Speaker Independent High resourced
7	VQ	Choosing any two dimensions, vector inspection, data plotting, checking whether data region for two different speakers are overlapping each other and in the same cluster, observation is needed, training of the vector VQ	It is suitable for losing data compression. Identification of high dimensional data. Helps to produce highly discriminative reference vectors for classifying state patterns. Easy to compute. Requires little storage.	It is too simple to act as an acoustic model in modern LVSCR system. It is text dependent because it requires codebook for matching.	Isolated Speech Recognition

		codebook using LGB algorithm function (vqlbg)			
8	SVM	Transformation of input data to higher dimensional space and then construction of linear binary into the high dimensional plane.	It helps in classification of speech data. Capability of dealing with data of higher dimensionality	It doesn't have convergence and stability problems	Isolated speech recognition. Continuous speech recognition. Speaker dependent.
9	ANN	Initialize the weight to 0, compute the output and then update the weights	Static patterns classification including noisy and acoustic data. Used to achieve human-like performance in ASR application. It can handle noisy low quality data efficiency. It requires minimum training of data vocabulary. It is easy to use and understand compared to statistical approaches.	It is very expensive because the training requires much iteration over large vocabulary data sets. Requires more training time More variation occurs due to neural network complexity.	Pure ANN is for Isolated word recognition Connected word recognition Speaker independent High resourced language.

In Table 2, it can be deduced that GMM works better for under-resourced languages compared to other approaches. The application areas expatiates where each approaches performs ultimately well for each classification mode. This gives an avenue to create a hybrid, try-bid or more models together for speech recognition enhancement.

6 SPEECH RECOGNITION TOOLS

Praat. It is a free computer software developed by Paul Boersma and David Weenik of University of Amsterdam. It runs on different variety of operating system platforms starting from Unix, Linux, Mac and Microsoft Windows. It offer a wide range of both standard and non-standard procedures ranging from spectrographic analysis, neural networks and articulatory synthesis. It is an open source code software developed by academician. Its default language support is in English, default programming language is C, C++. Its Strength includes acoustic analysis, manipulation, synthesizing and annotation of speech for both individual and multiple sounds with extensive help function.

Audacity. It is a free, cross platform computer software developed by Dominic Mazzoni and Roger Dannenberg at Carnie Mellon University. It runs on different variety of operating system platforms starting from Microsoft Windows, macOS, Linux and Unix-like systems. It is a closed source code software developed by academician. Its default language support is of 36 different languages ranging from Afrikaans, Arabic, English, Dutch, etc., default programming language is ANSI C, and description is HMM Neural net. Its strength includes recording audio, post-processing of audio, editing, and multi-track mixing. Its drawback is that it only supports 32-bit or64 bit VST audio effect plugins which depends on the architecture it was built for, lacks dynamic equalizer controls and real rime effects to support recording and also doesn't support restricted file formats wither AAC, AA3, or WMA.

Cmu Sphinx. It is a free, cross platform speaker independent large vocabulary continuous speech recognizer developed at Carnegie Mellon University. It comprises of different packages like sphinx-base, pocket-sphinx, sphinx-train and sphinx4. It runs on different variety of operating system platforms starting from Microsoft Windows, macOS, Linux. It is a closed source code software developed by academician. Its default language support is of English (US, UK), German, Mandarin, Russian, French, default programming language is JAVA, C, description is HMM, GMM. Its strength includes supporting low-resourced languages, free access to different language and acoustic models. Its drawback includes necessity for large database or the recognition accuracy will be modest.

Julius. It is a free, two-pass large vocabulary continuous speech recognizer developed by Lee Akinobu at Kyoto University. It runs on different variety of operating system platforms starting from Microsoft Windows, Unix and Linux. It

is a closed source code software developed by academician. Its default language support is of Japanese and English, default programming language is C, description is HMM trigrams, context-dependent HMM, Its strength includes performing real time computing decoding on most current personal computers. Its drawback includes supporting only two languages which are English and Japanese.

Dragon. It is a discrete speech recognition computer software developed by Dr. James Baker of Nuance Communications. It runs on different variety of operating system platforms starting

from Microsoft Windows and macOS. It is an open source code software developed by industry. Its default language support is of 8 different languages which are English (UK, US), French, German, Italian, Spanish, Dutch and Japanese. Its description is HMM and temporal pattern recognition. Its strength includes suitability for recognition tasks due to simple API code and god documentation. Its drawback includes not supporting background windows dictation, inability to determine boundaries of words during continuous speech input, limited free functionality of not more than 10000 request per day and above 10000 for paid access, complex licensing system and difficulty in implementing custom product.

Google Api. It is a free, cross platform speaker independent large vocabulary continuous speech recognizer developed at Carnegie Mellon University.. It runs on different variety of operating system platforms starting from Microsoft Windows, MacOS, Linux. It is an open source code software developed by academician. Its default language support is of 120 languages, default programming language is C, and description is Neural Networks. Its strength includes supporting customization, work in both real time and batch modes, fast, more embeddable, noise robustness, and speaker diarization. Its drawback includes supporting limitation of 60 minutes per day and more than 60 minutes for paid access.

Siri. It is an intelligent virtual assistant developed by Apple. It runs on variety of operating system platforms starting from iOS 5, macOS, tvOS, watchOS, and ipadOS. It is an open source code software developed by the industry. Its default language support is of English, French, Japanese, Dutch, Cantonese, Turkish, etc. which are 21 in total. Its strength includes supporting wide range of user commands, navigation and engaging with IOS integrated applications. Its drawback includes requirement of stiff user commands, lack of flexibility, inability to understand certain English accents.

Yandex Speech. It is also a speech recognition computer software developed by Yandex. It runs on variety of operating system platforms starting from Microsoft Windows, macOS, and Linux. It is an open source code software developed by the industry, default language support is of Russian, English and Turkish, Description include RNN (recurrent neural network). Its strength includes supporting computer games and applications, interactive voice response. Its drawback includes limitation of not more than 10000 request per day.

Microsoft Speech API. It is a free speech recognition computer software developed by Microsoft. It only runs on different variety of Microsoft Windows operating system platform. It is an open source code software developed by industry. Its default language support is of 29 languages, default programming language is C, description include Context dependent deep neural hidden markov model (CD-DNN-HMM). Its strength includes real time processing, customization, text normalization and profanity filtering. Its drawback include supporting up to 5000 request per month and more for paid access.

Tensorflow. It is a free speech recognition computer software developed by Google Brain Team which is a library for multidimensional array operation efficiency and it is based on deep learning. It only runs on Android, macOS, Windows, Android, and Linux operating system platform. It is an open source code software developed by industry. Its default language support is several, default programming language is Python, C++, CUDA and Java. Description include Neural Networks. Its strength includes GPU, TPU and CPU support which are efficient in performing highly parallelized numerical computations, pipelining, seamless/high performance, serves as a visualization and monitoring tool. Its drawback includes unique structure making debugging difficult.

Ibm Watson. It is a speech recognition computer software developed by Thomas Watson. It only runs on different variety of Microsoft Windows operating system platform. It is closed source code software developed by industry. Its default language support is of 8 languages. Its strength includes pronunciation customization, expressiveness, and custom words, speaker diarization. Its drawback includes supporting 1 million characters per month and more than for paid access.

KALDI. It is a free speech recognition computer software developed by Daniel Povey and others. It runs on different variety of operating system platforms starting from Unix and Microsoft Windows is an open source code software developed by academician, the default language support is only English, default programming language is C++, description is GMM, Neural Nets, Its strength includes flexibility, extensibility, cleanly structured code. Its drawback include documentation is only oriented to experience readers and not beginners.

7 APPLICATION OF SPEECH SYNTHESIS

Applications for the Vocally and Visually Impaired. Utilization of a handheld, battery powered synthetic speech aid can be done by vocally handicapped person well efficent word expression having specially designed keyboard that accepts input, and then converts it into speech immediately and for the visually impaired, the blind rely on the ability to hear and listen to information through tape or CD. Applications for Telecommunication and Multimedia. The speech synthesis technology gives the possibility to access information vocally with the aid of a telephone where queries are put through using user's voice or telephone keyboard. Applications in Education and Games. Speech synthesis can be used by many education institution or sports. If a lecturer is tired, he/she can decide to make use of a speech synthesizer computer to lecture the whole day with the same performance, efficiency and accuracy. Voice Enabled Email. Speech Synthesizer can also be used in voice enabled email to access their email from any telephones/smartphones where a user can dials a phone number to access a voice portal like saying a phrase, "Get my e-mail". They are useful for mobile workers to access their email easily from virtually anywhere. Human-Machine interactions, Hands free computing, Interactive voice response, and Home automation, voice-enabled email. Speech synthesizer can also be used many kinds of human to machine interaction and different interfaces which includes alarm systems, clocks, washing machines, etc. [10] [21]. Application in Transportation, Medical and Entertainment. Speech synthesizer can also be used in hospitals system, airlines, and entertainment world [22].

8 DISCUSSION AND CONCLUSION

Review of speech recognition concepts has been done from earliest study till date involving the techniques, tools, approaches and algorithms. These led to highlighting their steps, limitations/weaknesses, strengths, applications areas etc. As of today, speech recognition models should prove to be more ideal and fit especially for under-resourced languages which is critically needed for any future application development. As was presented, this current study did not include hypothesis in its model formulation and testing but rather gives a comparative analysis of speech recognition tools, approaches and techniques [27]. In future research, health related ethical issues may be cinsidered [28].

9 ACKNOWLEDGMENT

In this study, we would like to express our deep appreciation for the support provided by Covenant University Centre for Research, Innovation and Discovery (CUCRID).

REFERENCES

[1]. Besacier, L., & Barnard, E. (2014). Automatic Speech Recognition for Under-Resourced Languages : A Survey. (January 2018). https://doi.org/10.1016/j.specom.2013.07.008

[2]. Choudhary, A., Chauhan, R. S., & Gupta, G. (2013). Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK).

[3]. Martha, T., Solomon, Y., Abate, T., & Besacier, L. (2013). Using different acoustic, lexical and language modeling units for ASR of an under-resourced language – Amharic", Speech Communication.

[4]. Das, P., Acharjee, K., Das, P., & Prasad, V. (2016). VOICE RECOGNITION SYSTEM : SPEECH-TO-TEXT. (July).

[5]. Kozierski, P., Sadalla, T., Drgas, S., & Giernacki, W. (2018). Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition. 16, 109–114. https://doi.org/10.15439/2018F255

[6]. Levis, J. M., & Suvorov, R. (2012). Automatic Speech Recognition. (July 2018). https://doi.org/10.1002/9781405198431.wbeal0066

[7]. Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A Review on Speech Recognition Technique. International Journal of Computer Applications, 10(3), 16–24.

[8]. Frederic, W., Pirates, T., Lost, P., & James, H. (1998). Chapter 9 Automatic Speech Recognition. 285–334.

[9]. K.Saksamudre, S., Shrishrimal, P. P., & Deshmukh, R. R. (2015). A Review on Different Approaches for Speech Recognition System. International Journal of Computer Applications, 115(22), 23–28. https://doi.org/10.5120/20284-2839

[10]. Khilari, P., & P, P. B. V. (2015). Implementation of Speech to Text Conversion. 6441–6450. https://doi.org/10.15680/IJIRSET.2015.0407167

[11]. Lööf, J., Gollan, C., & Ney, H. (2009). Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 88–91.

[12]. Nguyen, M., Vo, T., (2015). Vietnamese Voice Recognition for Home Automation using MFCC and DTW Techniques. International Conference on Advanced Computing and Applications (ACOMP).

[13]. Srinivasan, A. (2013). Real time speaker recognition of letter 'zha' in Tamil language. Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).

[14]. Ltaief, A. Ben, Estève, Y., Graja, M., & Belguith, L. H. (2017). Automatic speech recognition for Tunisian dialect. 1–8.

[15]. Adetunmbi, O. A., Obe, O. O., Iyanda, J. N. Development of Standard Yorùbá speech-to text system using HTK. International Journal of Speech Technology, 2016.

[16]. Abhishek Reddy, K. N., Agrawal, P., Singh, P., Singh, P., & N.R, L. (2017). A Comparative Study on Speech Recognition Approaches and Models. International Journal of Computer Trends and Technology, 43(2), 118–123. https://doi.org/10.14445/22312803/ijctt-v43p117

[17]. Khilari, P., & P, P. B. V. (2015). Implementation of Speech to Text Conversion. 6441–6450. https://doi.org/10.15680/IJIRSET.2015.0407167

[18]. Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A Review on Speech Recognition Technique. International Journal of Computer Applications, 10(3), 16–24. https://doi.org/10.5120/1462-1976

[19]. Kumar, J., Prabhakar, O. P., & Sahu, N. K. (2014). Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems : A Review. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), 2760–2769.

[20]. Haton, J. P. (2003). Automatic speech recognition: A Review. ICEIS 2003 - Proceedings of the 5th International Conference on Enterprise Information Systems, 1(9), IS5–IS10. https://doi.org/10.5120/9722-4190

[21]. Nanayakkara, L., Liyanage, C., Tharaka Viswakula, P., Nagungodage, T., Pushpananda, R., & Weerasinghe, R. (2018). A Human Quality Text to Speech System for Sinhala. (February 2019), 157–161. https://doi.org/10.21437/sltu.2018-33

[22]. Ajayi, L. K., Azeta, A. A., Owolabi, I. T., Damilola, O. O., Chidozie, F., Azeta, A. E., & Amosu, O. (2019, August). Current Trends in Workflow Mining. In Journal of Physics: Conference Series (Vol. 1299, No. 1, p. 012036). IOP Publishing.

[23]. Trivedi, P. A. (2014). Introduction to Various Algorithms of Speech Recognition : Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks. International Journal of Engineering Development and Research, 2(4), 3590–3596.

[24]. Anusuya, M. A., & Katti, S. K. (2009). [2009] Speech Recognition by Machine: A Review. IJCSIS) International Journal of Computer Science and Information Security, 6(3), 181–205.Retrievedfromhttp://sites.google.com/site/ijcsis/

[25]. Ghai, W., & Singh, N. (2012). Literature Review on Automatic Speech Recognition. International Journal

of Computer Applications, 41(8), 42–50. https://doi.org/10.5120/5565-7646

[26]. Rami, M., Svitlana, M., Lyashenko, V., & Belova, N. (2017). Speech Recognition Systems : A Comparative Review. IOSR Journal of Computer Engineering, 19(5), 71–79. https://doi.org/10.9790/0661-1905047179

[27]. [27]Azeta A. A., Ayo C. K., Atayero A. A. and Ikhu-Omoregbe N. A. (2009), A Case-Based Reasoning Approach for Speech-Enabled e-Learning System", 2nd IEEE International Conference on Adaptive Science & Technology (ICAST). December 14 – 16, 2009, Accra Ghana. ISBN: 978-1-4244-3523-4, ISSN: 0855-8906. PP. 211-217. Index in Scopus. Available online at IEEE Xplore database tab 4: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5375737

[28]. . [28] Victor C. Osamor, Ambrose A. Azeta and Oluseyi O. Ajulo (2013), "Tuberculosis–Diagnostic Expert System: An architecture for translating patients information from the web for use in tuberculosis diagnosis. SAGE Journal. Health Informatics Journal 19 (3). (ISI Journal impact factor 0.8). Index in Scopus and Thomson Reuters Web of Science.