SPEECH RECOGNITION TECHNIQUES ON MOBILE DEVICES:- ANALYSIS OF VARIOUS APPROACHES

[1]GULBAKSHEE DHARMALE, [2]DR. DIPTI D. PATIL, [3]DR. VILAS THAKARE

[1]Research Scholar, [1,3]SGB Amravati University, Amravati, INDIA,
[2]MKSSS's CCOEW, SavitribaiPhule, Pune University, Pune, INDIA

Abstract- Speech is the most common and convenient way to communication. Speech is also faster than typing on a keypad and more expressive than clicking on a menu item. For these reason Automatic Speech Recognition (ASR) is become very important and popular in today's world. Speech recognition is the process of converting spoken words into text. After years of research and development the accuracy of automatic speech recognition remains one of the important research challenges (e.g. variations of the context, speakers, and environment).The design of speech recognition system requires careful attentions to different issues such as: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, performance evaluation and database. One of the major problems faced in speech recognition is that the spoken word can be vastly altered by accents, dialects and mannerisms. This paper presents review of Automatic Speech Recognition techniques with Artificial Intelligence related to mobile platform.

Keywords- Automatic Speech Recognition, Speech recognition Techniques, Feature Extraction, Artificial Intelligence.

I.INTRODUCTION

Speech is a tool of communication, also a symbol of identity and authorization. The idea of speech-based recognition comes from the human imagination and creativity that have been frequently used in television programs and several movies. Speech recognition-based authentication has been presented as the symbol of technological advancement as well as a secure system.

Smartphone's and tablets are rapidly overtaking desktop and laptop, computers as people's primary computing device. They are heavily used to access the web, read and write messages, interaction social networks, etc. [1]. This popularity comes despite thefact that it is significantly more difficult to input text on these devices, mostly by using an on-screen keyboard. Automatic speech recognition is alternative to typing on mobile services. It is a natural and increasingly popular. Google offers the ability to search by voice [2] on Android, iOS and Chrome; Apple's iOS devices come with Siri, a conversational assistant. On both Android and iOS devices, users can also speak to fill in any text field where they can type, a capability greatly used to dictate SMS messages and e-mail [3].

Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop systems and techniques for speech input to machine. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert it to a machinereadable format. In today's world many speech recognition applications, such as voice dialing, simple data entry and speech-to-text are exist. While speech recognition sets its goals at recognizing the spoken words in speech, the main aim of speaker recognition is to identity the speaker by characterization, extraction and recognition of the information confined in the speech signal [4].

As these mobile devices are often used when the person is "on the move", variable acoustic environments and limited resources on the mobile device needs special arrangements. The Automatic Speech Recognition is a software technology that allows a machine to extract the message, oral contained in a speech signal [5]. This technology uses computational methods in areas of signal processing and artificial intelligence.

Fig.1 shows a mathematical representation of speech recognition system in simple equations which contain front end unit, model unit, language model unit, and search unit. The recognition process is shown below (Fig .1).

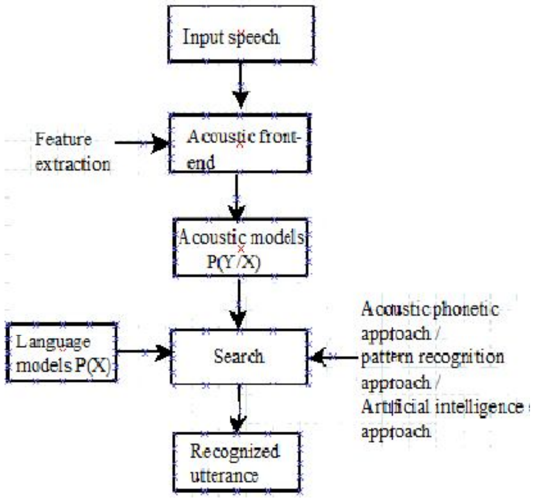


Fig.1. Basic model of speech recognition

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence X, produces an acoustic observation sequence Y, with probability P (X, Y). The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability.

$$P\left(\frac{X}{Y}\right) = \arg max_X P\left(\frac{X}{Y}\right) \qquad -\ (1)$$

Using Bay's rule, equation (1) can be written as

$$P\left(\frac{X}{Y}\right) = \frac{P\left({}^{Y}/{}_{X}\right)P(X)}{P(Y)} \qquad -\ (2)$$

Since P(Y) is independent of X, the MAP decoding ruleofequation(1) is

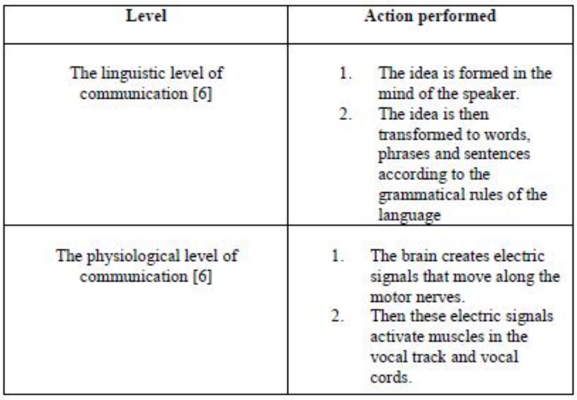
$$X = agr\ max_X P\left(\frac{Y}{X}\right)P(X) \qquad -\ (3)$$

The first term in equation (3) P (Y/X), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. Hence P(Y/X) is computed. It is necessary to build statistical models for sub word speech units, build up word models from these sub word speech unit models (using a lexicon to describe the composition of words) and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods, for large vocabulary speech recognition systems. The second term in equation (3) P(X), is called the language model. It describes the probability associated with a postulated sequence of words. Such language models can combine both syntactic and semantic constraints of the language and recognition task.

II. RELATED WORK

Gaussian mixture models (GMMs) is used by most conventional speaker recognition systems to capture framelevel characteristics of a person's voice, where the speech frames are assumed to be independent of one another as the physical shape of a human vocal tract is different from person by person. Hence, each human speaks in a different way. If a person is asked to utter the same word twice, the speech signal will not be exactly same as the frequency and other sound properties may differ from time to time. There are some features that makes difficult to understand speech signals these are; an environment where human speaks, the dialect of the language, differences in the vocal tract length of males, female and children provide the speech variation.Though, there are still some features in the human speech which can be mathematically modelled and used for predicting words from it but it demands tremendous amount of time and effort [6]. To model a human hearing system, it is important to understand the working of human auditory system which is shown in table 1 given below:

Table 1: Working of human auditory system



Level	Action performed
The linguistic level of communication [6]	1. The idea is formed in the mind of the speaker. 2. The idea is then transformed to words, phrases and sentences according to the grammatical rules of the language
The physiological level of communication [6]	1. The brain creates electric signals that move along the motor nerves. 2. Then these electric signals activate muscles in the vocal track and vocal cords.

Because of this independence assumption, GMMs often fail to capture certain types of speaker-specific information that evolve over time scales of more than one frame. For example, since words usually span many frames, GMMs [7] tend to be poorly suited for modelling differences in word usage (idiolect) between speakers. In recent times, automatic speaker recognition research has expanded from utilizing only the acoustic content of speech to examining the use of higher levels of speech information, commonly referred to as "highlevel features." Anencouraging direction in high-level feature research has been the use of n-gram based models to capture speaker specific patterns in the phonetic and lexical content of speech.

In, Doddington performed an important study about using the lexical content of speech for speaker recognition and an ngram based technique is introduced for modelling a speaker's idiolect. This trend in research was continued by Andrews, Kohler, and Campbell among others, who used similar n gram, based models to capture speaker pronunciation idiosyncrasies through analysis of automatically recognized phonetic events. This line of research is generally referred to as "Phonetic Speaker Recognition." The research of Andrews et al. and Doddington showed word and phone n-gram based models to be quite promising for speaker recognition.

There have been numerous attempts, especially since the Johns Hopkins 2002 Workshop to harness the power of all kinds of high-level features. The relative frequencies of phone n-grams as features for training speaker models and for scoring test-target pairs used by the current "state-of-the-art" in phonetic speaker recognition[8]. Typically, these relative frequencies are computed from a simple 1-best phone decoding of the input speech. The phonetic speaker recognition research work has been extended in various ways by introducing different modelling strategies and different methods of utilizing the source information such as described in Navratil. It proposed a method involving binary-treestructured statistical models for extending the phonetic context beyond that of standard n-gram (particularly bigrams) by exploiting statistical dependencies within a longer sequence window without exponentially increasing the model complexity, as is the case with n-grams. The described approach confirms the relevance of long phonetic context in phonetic speaker recognition and represents an intermediate stage between short phone context and word-level modeling without the need for any lexical knowledge. Binary-tree model represents a step towards flexible context structuring and extension in phonetic speaker recognition, consistently outperforming standard smoothed bigrams as well as trigrams.

A conditional pronunciation modelling method is proposed by Klusacek. It uses time-aligned streams of phones and phonemes to model a speaker’s specific pronunciation. The system uses phonemes drawn from a lexicon of pronunciations of words recognized by an automatic speech recognition system to generate the phoneme stream and an open-loop phone recognizer to generate a phone stream. The phoneme and phone streams are aligned at the frame leveland conditional probabilities of a phone, given phoneme, are estimated using co-occurrence counts. A probability detector is then applied to these probabilities for the speaker detection task. This approach achieves a relatively high accuracy in comparison with other phonetic methods in the Super SID project at the Johns Hopkins 2002 Workshop.

III. EFFECTIVE TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION

Basically there exist three techniques to speech recognition.
These are:

1. Acoustic Phonetic Approach [9]
2. Pattern Recognition Approach [10]
3. Artificial Intelligence Approach [5, 9]

Classification of these speech recognition techniques shown in given tree diagram.

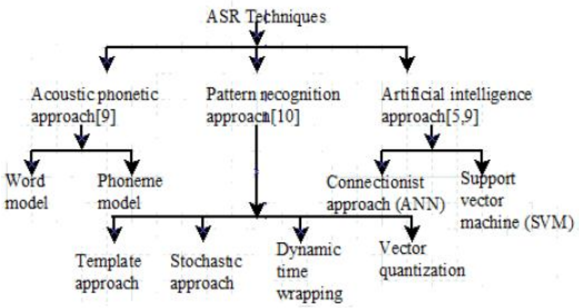

ASR Techniques
Acoustic phonetic approach[9]
Pattern recognition approach[10]
Artificial intelligence approach[5,9]
Word model
Phoneme model
Connectionist approach (ANN)
Support vector machine (SVM)
Template approach
Stochastic approach
Dynamic time wrapping
Vector quantization

Fig. 2. Automatic Speech Recognition Approaches

3.1 Acoustic Phonetic Approach

Acoustic phonetic approachis the oldestapproach to speech recognition were based on finding speech sounds and allocatecategorized labels to these sounds which is the basis of the acoustic phonetic approach, which postulates that there exist finite, distinctive phonetic unitsin spoken language and these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. There are three basic steps for the acoustic phonetic approach. These are as follows [9]:

The first step is a spectral analysis which combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units.

The second step is speech segmentation and labeling phase in this phase the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech.

The third and last step in this approach is validation phase which attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labelling. In the validation process, linguistic constraints on the task (i.e., the vocabulary, the syntax, and other semantic rules) are invoked in order to access the lexicon for word decoding based on the phoneme lattice.

There are two types of acoustic models i.e. word model and phoneme model. An acoustic model is implemented using different approaches such as HMM, ANNs, Dynamic Bayesian Networks (DBN), Support Vector Machines (SVM). HMM is used in some form or the other in every state of the art speech and speech recognition system [9].

3.2 Pattern Recognition Approach

Pattern trainingand pattern comparison are involves in the pattern-matching approach. The fundamental feature of pattern-matching approach is that it uses awell formulatedmathematicalframework and creates consistent speech pattern representationsfor reliable pattern comparison from a set of labeled training samples via a formal trainingalgorithm. A speech pattern representation can be in the form of a speech templateor a statistical model and can be applied to a sound (smaller than a word), a word ora phrase. In the pattern-comparison stage of the approach, a direct comparison is madebetween the unknown speeches (the speech which needs to be recognized) with each possible patternlearned in the training stage in order to determine the identity of the unknown accordingto the goodness of match of the patterns. Usually, pattern recognition approachesare model based such as Hidden Markov Model (HMM), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Vector Quantization (VQ) and Dynamic TimeWarping (DTW) [10].

3.2.1 Template Approach

A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate's words. Recognition is then carried out by matching an unknown spoken utterance with each of these references.

There are two main ideas in template method these are:

1. One key idea is to derive typical sequences of speech frames for a pattern (a word) via some averaging procedure and to depend on the use of local spectral distance measures to compare patterns.
2. Another key idea is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across speakers as well as across repetitions of the word by the same speaker [9].

3.2.2 Stochastic Approach

Stochastic modelling entails the use of probabilistic models to deal with incomplete or uncertain information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, contextual effects, speaker variability, confusable sounds and homophones words [9].

A common issue in pattern recognitionis high dimensionality of feature vectors. There are many reasons for having high dimensional feature spaces. For instance, static features can be augmented with dynamic spectral information in speech feature extraction. One way of achieving this is by combining multiple consecutive Mel-filtered cepstrum coefficients (MFCC) feature vectors to form high dimensional super-vectors that may represent on the order of 100 milliseconds of speech. These super feature vectors can have very high dimensionality, which may lead to significant problems when performing a pattern recognition task. Therefore, it is a good practice to perform some sort of dimensionality reduction before applying a particular pattern recognition algorithm to these features. Intuitively, a good dimensionality reduction algorithm should be able to preserve important information from the original feature space in the low dimensional transformed feature vectors [11].

3.2.3 Dynamic Time Warping (DTW)

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. DTW is a well-known application of automatic speech recognition used to deal with different speaking speeds. In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g., time series) with certain restrictions. That is, the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models [12].

3.2.4 Vector Quantization (VQ)

The objective of VQ is theformation of clusters, each representing a specific class. During the training process, extracted feature vectors fromeach specific class are used to form a codebook, throughthe use of an iterative method. Thus, the resulting codebookis a collection of possible feature vector representations foreach class. During the recognition process, the VQalgorithm will go through the whole codebook in order tofind the corresponding vector, which best represents theinput feature vector, according to a predefined distancemeasure[13].

3.3 Artificial Intelligence Approach (Knowledge Based Approach)

Artificial intelligence approach is a hybrid of the acoustic phoneticapproach and pattern recognition approach. In this, it exploits the concepts and ideasof acoustic phonetic and pattern recognition methods. Knowledge based approach usesthe information regarding linguistic, phonetic and spectrogram [10]. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; it provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult [9]. There are different techniques and algorithms comes under Artificial Intelligence approach.

3.3.1 Connectionist Approaches (Artificial Neural Networks)

The artificial intelligence approach try to automate the recognition procedure according to the way a person applies intelligence in visualizing, analyzing, and characterizing speech based on a set of measured acoustic features. Among the techniques used within this class of methods are uses of an expert system (e.g., a neural network) that integrates phonemic, lexical, syntactic, semantic and even pragmatic knowledge for segmentation and labeling that uses tools such as artificial Neural Networks for learning the relationships among phonetic events. The focus in this approach has been generally in the representation of knowledge and integration of knowledge sources.

Connectionist modeling of speech is the earliest development in speech recognition and still the subject of much controversy. In connectionist models, knowledge or constraints are distributed across many simple computing units instead of encoded in individual units, rules or procedures. Uncertainty is modeled not as likelihoods or probability density functions of a single unit, but by the pattern of activity in many units. The computing units are simple in nature and knowledge is not programmed into any individual unit function; rather, it lies in the connections and interactions between linked processing elements. The uniformity and simplicity of the underlying processing element makes connectionist models attractive for hardware implementation, which enables the operation of a remaining to be simulated efficiently [9].

3.3.2 Support Vector Machine (SVM)

SVM is one of the most efficient machine learning algorithms, which is mostly used for pattern recognition. The basic idea of the SVMs is building an optimal hyper plane in order to use in classification of linearly separable patterns. An optimal hyper plane is a selected one from a set of hyper planes which maximizes hyper plane margin, which is the distance from hyper plane to the nearest point of the pattern. SVM is primarily set to maximize the margin, which will guaranty that the input pattern would be classified correctly. In order to classify data, based on either a priori information or statistical data mined from raw data set, pattern recognition is widely utilized and this makes it an extremely powerful tool for data separation in many fields. The support vector machine (SVM) usually copes with pattern classification that means this algorithm is used mostly for classifying the different types of patterns. Now, there are two different type of patterns such as Linear and non-linear. Linear patterns areeasily distinguishable or can be easily separated in low dimension whereas non-linear patterns are not easily distinguishable or cannot be easily separated and hence these type of patterns need to be further manipulated so that it can be easily separated [14].

IV. DIFFERENT SPEECH RECOGNITION BLOCKS

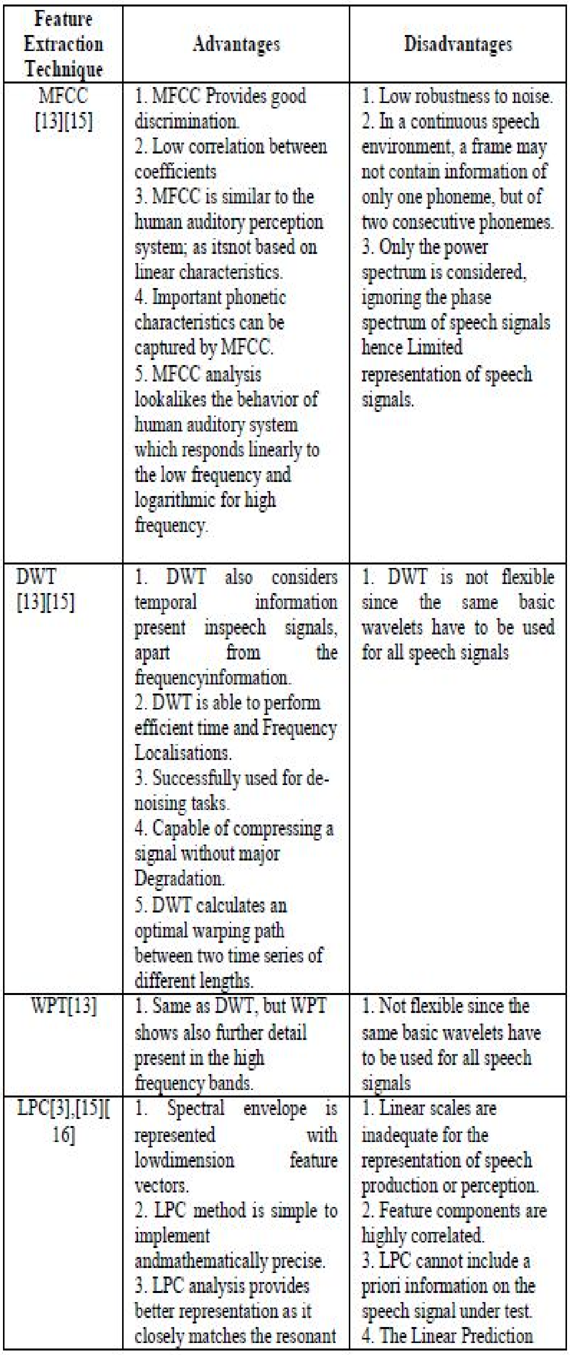
A general speech recognition system consists of four blocks:

- Feature extraction, Language modelling, Pronunciation modelling, Decoder. This blocks help to simplify the face recognition models.

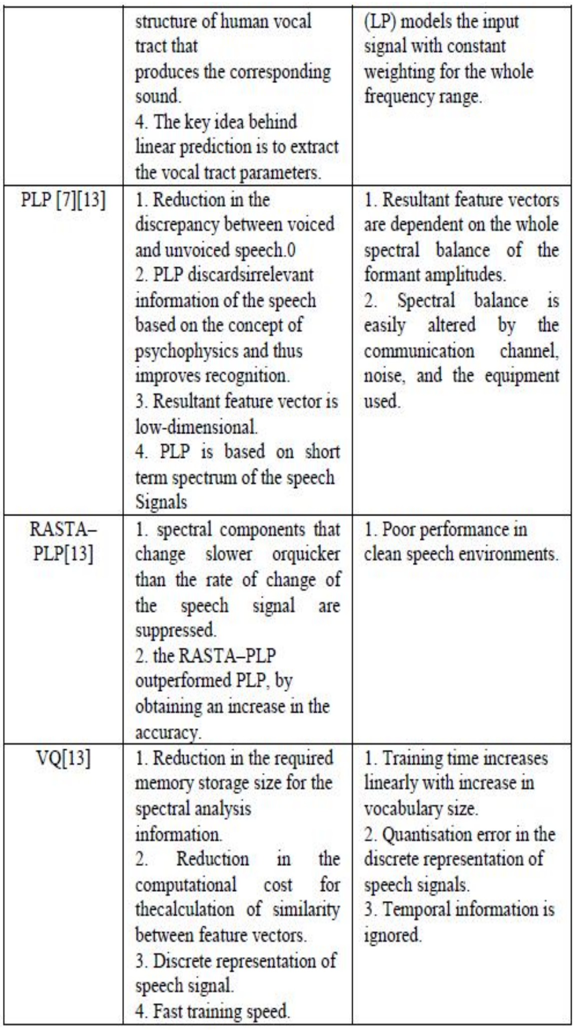
4.1 Feature Extraction

In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. Passing huge speech data as inputs to a machine learning algorithm is an exhaustive task with no extra merits for the size of data. Large data set might be an obstacle for enrollment and training stages which may prevent the system from learning. It greatly affects the system performance. Due to that, the input data set should be transformed into a reduced representation set of features this transformation process is called feature extraction. The features should be representative and carefully selected which leads to better and faster recognition [7]. The process of feature extraction reduces speech data while maintaining the discriminative information of a speaker voice signal [10]. There are different techniques available for the feature extraction of speech like MFCC, RAST, PLP, LPCC, but MFCC is the most commonly used feature extraction technique for speech[5]. Table 2 shows the advantages and disadvantages of different features extraction techniques used for the voice recognition.

Table 2.Advantages and disadvantages of feature extractiontechniques [13].



Feature Extraction Technique	Advantages	Disadvantages
MFCC [13][15]	1. MFCC Provides good discrimination. 2. Low correlation between coefficients 3. MFCC is similar to the human auditory perception system; as itsnot based on linear characteristics. 4. Important phonetic characteristics can be captured by MFCC. 5. MFCC analysis lookalikes the behavior of human auditory system which responds linearly to the low frequency and logarithmic for high frequency.	1. Low robustness to noise. 2. In a continuous speech environment, a frame may not contain information of only one phoneme, but of two consecutive phonemes. 3. Only the power spectrum is considered, ignoring the phase spectrum of speech signals hence Limited representation of speech signals.
DWT [13][15]	1. DWT also considers temporal information present inspeech signals, apart from the frequencyinformation. 2. DWT is able to perform efficient time and Frequency Localisations. 3. Successfully used for de-noising tasks. 4. Capable of compressing a signal without major Degradation. 5. DWT calculates an optimal warping path between two time series of different lengths.	1. DWT is not flexible since the same basic wavelets have to be used for all speech signals
WPT[13]	1. Same as DWT, but WPT shows also further detail present in the high frequency bands.	1. Not flexible since the same basic wavelets have to be used for all speech signals
LPC[3],[15][16]	1. Spectral envelope is represented with lowdimension feature vectors. 2. LPC method is simple to implement andmathematically precise. 3. LPC analysis provides better representation as it closely matches the resonant	1. Linear scales are inadequate for the representation of speech production or perception. 2. Feature components are highly correlated. 3. LPC cannot include a priori information on the speech signal under test. 4. The Linear Prediction



	structure of human vocal tract that produces the corresponding sound. 4. The key idea behind linear prediction is to extract the vocal tract parameters.	(LP) models the input signal with constant weighting for the whole frequency range.
PLP [7][13]	1. Reduction in the discrepancy between voiced and unvoiced speech.0 2. PLP discardsirrelevant information of the speech based on the concept of psychophysics and thus improves recognition. 3. Resultant feature vector is low-dimensional. 4. PLP is based on short term spectrum of the speech Signals	1. Resultant feature vectors are dependent on the whole spectral balance of the formant amplitudes. 2. Spectral balance is easily altered by the communication channel, noise, and the equipment used.
RASTA–PLP[13]	1. spectral components that change slower orquicker than the rate of change of the speech signal are suppressed. 2. the RASTA–PLP outperformed PLP, by obtaining an increase in the accuracy.	1. Poor performance in clean speech environments.
VQ[13]	1. Reduction in the required memory storage size for the spectral analysis information. 2. Reduction in the computational cost for thecalculation of similarity between feature vectors. 3. Discrete representation of speech signal. 4. Fast training speed.	1. Training time increases linearly with increase in vocabulary size. 2. Quantisation error in the discrete representation of speech signals. 3. Temporal information is ignored.

4.2 Decoder

Decoding is the most essential step in the speech recognition process. Decoding is performed for finding the best match for the incoming feature vectors using the knowledge base [5]. The objective of the decoder is to find out the most probablesequence of words from the language model, produces theobservation sequence[17].

The Viterbi beam search algorithm is used by the decoding stage to find the most likely sequence of phones for the observed speech. Each stage in the algorithm uses models, which represent the probabilities of sounds, sequences of sounds, words and sequences of words in the language. Gaussian distributions are used to represent nature of the sounds and HMMs are used to model sequences and duration of the sounds. Word sequences and their probabilities are stored as weights, which are added during the decode process [17].

4.3 Language Modeling

Language models are used to guide the search correct word sequence by predicting the possibility of nth word using (n-1) preceding words. Language models can be classified into: [5]

1. Uniform model: In uniform model each word has equal probability of occurrence.
2. Stochastic model: In stochastic modelprobability of occurrence of a word depends on the word preceding it.
3. Finite state languages: languages use a finite state network to define the allowed word sequences.
4. Context free grammar: It can be used to encode which kind of sentences is allowed.

4.4 Pronunciation Modelling

In pronunciation modelling, during recognition the sequence of symbols generated by acoustic model HMM is compared with the set of words present in dictionary to produce sequence of words which is the system's final output contains information about which words are known to thesystem and how these words are pronounced i.e. what is their phonetic representation. Decoder is then used for recognizing words by combining and optimizing the information of acoustic & language models [5].

V. ADVANCED ASR BASED TECHNIQUES

There are different basic ASR techniques which are modified by different researchers for the speech recognition in the mobile devices is called advanced ASR based techniques. Some of the sear studied in this section.

5.1 Large Vocabulary Continuous Speech Recognition

Research and development activities in the area of Large Vocabulary Continuous Speech Recognition (LVCSR) are concentrated on developing a dictation system. There are also noise models trained for better modelling of the inter word and inter sentence noise which could produce false tri-phones detection. In the language modelling task during the last years, it has encountered several problems with text pre-processing, selection of the basic statistical methods used in the modeling of the other languages and adaptation into the area of application. Another important part in the process has been optimization of the resultant model, which introduced phonetic and linguistic relations between words. These optimization steps have caused an improvement in recognition accuracy of the LVCSR system [18].

5.2 An Arabic Speaker Verification System

In research work, it developed and analysed an Arabic speaker verification system with good accuracy. The system is used in an access control application for mobile devices to prevent unauthorized users from gaining access to the device. The process of speaker verification consists of two main stages; enrolment stage and verification stage [19].

Enrolment stage: In the enrolment stage, a speaker S repeats a set pass phrase n times. The speech signals of the speaker are pre-processed and features are extracted. After that, the features are passed to the support vector machine (SVM) as the enrolment data. By training SVM, a speaker statistical model for S is created. A background model for imposters is created as well; which uses speech signals from a variety of unauthorized speakers.

Verification stage: During the verification stage, a speaker X says the pass phrase and claims to be speaker S. The claimed speaker model and background model are fetched to verify the claimed identity. Similarly, the spoken phase goes through the same phases in the enrolment stage.

5.3 Android Applications using Voice Controlled Commands

Assistive Technology (AT) is dedicated to increasing the independence and mobility of the persons with disabilities. Persons with quadriplegia are affected by limitations in physical and independence. In this project, voice recognition is explored as a template upon which the independence of persons with neuromuscular disorders can be expanded. For this reason, android applications were developed on a Smartphone to operate a television remote via Bluetooth exchange and PIC processing.

The hardware implements a Bluetooth modem, BlueSMiRF silver (WRL-10269), which receives the commands from the android application software that was developed for the Smartphone. The BlueSMiRF modem communicates with the PIC ®Microcontroller (PIC18f4525- I/P) via the EUSART transmit and receive pins on the microcontroller, pins 25 and 26, respectively. Each function in the android application recognizes keywords which then send a specific signal via Bluetooth connection to the microcontroller. A simple code in the microcontroller interprets the signals received from the modem and sends a corresponding signal out of one of the seven outputs of port B to the quad bilateral switches. In order to trigger specific buttons on the direct TV remote control, reverse engineering was used by soldering wires from a ribbon cable to the specific button locations inside the remote [20].

5.4 Augmented Reality

Augmented reality enhances user perception by supplementing the real world through the additional of virtual content which is stored in a library. Augmented reality enables users to complement their actual environment with that of a simulated environment to provide enhanced information and data without sacrificing the information stored within the real environment. Augmented reality is used to describe a system that superimposes computer generated information overlaying the real environment. The goal of augmented reality is to superimpose informationin the form of audio, text, graphics and other sense enhancements over a real environment in real time.

Upon viewing potential mobile devices to serve as training tools, augmented reality is selected as a potential mobile technology. Augmented reality is selected here because it offers hands free operation through a head mounted display HMD. System is capable of transferring data between the associate and system without hand operation or relocation to a computer. User is able to control the device through audible commands and incorporates a simple and easy to use operator interface. The key benefit of this implementation is it will minimize, or potentially eliminate, the need for a training associate. A cost savings could be directly associated with implementation as a trainer will not be neededfor extended periods of time [21].

VI. ARTIFICIAL INTELLINGENCE BASED SPEECH RECOGNITION

The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing and finally making a decision on the measured acoustic features. Expert system is usedwidely in this approach (Mori et al., 1987) [22].

The Artificial Intelligence approach [23] is a fusion of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; It provided little insight about human speech processing, thus making error analysis and knowledge based system enhancement difficult. On the other hand, a large body of linguistic and phonetic literature provided insights and understanding to human speech processing [24].

In its pure form, knowledge engineering design involves the direct and explicitincorporation of expert's speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of manylevels of human knowledge phonetics, lexicalaccess, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the modelsand algorithms of other techniques such as template matching andstochastic modeling. This form of knowledge application makesan important distinction between knowledge and algorithms enable us to solve problems. Knowledge enables thealgorithms to work better. This form of knowledge based systemenhancement has contributed considerably to the design of allsuccessful strategies reported. It plays an important role in theselection of a suitable input representation, the definition of unitsof speech, or the design of the recognition algorithm itself.

CONCLUSION

Speech is the most convenient means of communication between people. Becauseof the technological curiosity to build machines that mimic humans or desire to automate work with machines like mobile phones, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades.In this paper the detail review on Speech recognition with Artificial Intelligence related to mobile device is done. The main aim of this paper is to study the Speech recognitiontechniques and how the artificial intelligence can be used for speech recognition purpose.

REFERENCES

[1] Anuj Kumar, AnujTewari, Seth Horrigan, Matthew Kam1, Florian Metze and John Canny, "Rethinking Speech Recognition on Mobile Devices", IUI4DR, California, USA, February 13, 2011, pp.1-6.

[2] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Google search by voice: A case study in Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics", Springer, 2010, pp. 61–90.

[3] B.Ballinger,C. Allauzen, A. Gruenstein, and J. Schalkwyk, "Ondemand language model interpolation for mobile speech input," in Proc. Interspeech, 2010.

[4] SadaokiFurui, "50 years of Progress in speech and Speaker Recognition Research", ECTI Transactions on Computer and Information Technology,Vol.1. No.2, November 2005.

[5] PreetiSaini, ParneetKaur,"Automatic Speech Recognition: A Review", international Journal of Engineering Trends and Technology- Volume4, Issue2, ISSN: 2231-5381, 2013,pp.132-136.

[6] Shanthi Therese S., Chelpa Lingam," Review of Feature Extraction Techniques in Automatic Speech Recognition", International Journal of Scientific Engineering and Technology (ISSN: 2277-1581), Volume No.2, Issue No.6,1 June 2013, IJSET@2013,pp : 479-484.

[7] AbdulrahmanAlarifi, IssaAlkurtass," SVM based Arabic Speaker Verification System for Mobile Devices", International Conference on Information Technology and e-Services, 978-1-4673-1166-3/12/IEEE ©2012 .

[8] Nicholas Mulhern, Neil McCaffrey, Nicholas Beretta, Eugene Chabot, Ying Sun, "Designing Android Applications using Voice Controlled Commands For Hands free interaction with Common Household Devices", 39th Annual Northeast Bioengineering Conference, IEEE 2013, pp. 265-266.

[9] Sanjivani S. Bhabad ,Gajanan K. Kharate, "An Overview of Technical Progress in Speech Recognition", International Journal of Advanced Research in Computer Science and Software EngineeringResearch Department of E & TC, Pune university, India, Volume 3, Issue 3, March 2013, pp. 488-497.

[10] ShufeiDuan · Jinglan Zhang · Paul Roe · Michael Towsey, "A survey of tagging techniques for music, speechand environmental sound",ArtifIntell Rev DOI 10.1007/s10462-012-9362-y © Springer Science+Business Media Dordrecht 2012.

[11] Vikrant Singh Tomar, "A Family of Discriminative Manifold Learning Algorithms and Their Application To Speech Recognition", IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 1, January 2014,pp.161-171.

[12] Parwinder pal Singh Er. Bhupindersingh, "Speech Recognition as Emerging Revolutionary Technology" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012.

[13] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, "Comparative study of automatic speech recognition techniques",IET Signal Process, Vol. 7, Issue 1, 2013, pp. 25–46.

[14] SasanKaramizadeh, Shahidan M. Abdullah , MehranHalimi , JafarShayan and Mohammad javadRajabi, "Advantage and Drawback of Support Vector Machine Functionality", International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 - 4, 2014 - Langkawi, Kedah, Malaysia.

[15] Sharada C. Sajjan, Vijaya C, "Comparison of DTW and HMM for Isolated WordRecognition", International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 978-1-4673-1039-0/12/$31.00 ©2012 IEEE

[16] Nagsen S. Bansod1, Siddharth B. Dadhade2, "Speaker Recognition using Marathi (Varhadi) Language", 978-1-4799-3966- 4/14 $31.00 © 2014 IEEE DOI 10.1109/ICICA.2014.

[17] Ojas A. Bapat, P. D. Franzon. R. M. Fastow, "A Generic and Scalable Architecture for a Large Acoustic Model and Large Vocabulary Speech Recognition Accelerator Using Logic on Memory", IEEE transactions on very large scale integration (VLSI) systems, 2014,pp.1-12 .

[18] Deng and D. Yu, "Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition", in Proc.ICASSP, 2007,pp. 445–448.

[19] Bilmes, J. ; Das, S. ; Duta, N, "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop" International Conference on Acoustics, Speech, and Signal Processing Proceedings. (ICASSP '03) IEEE 2003.

[20] C.H.Lee, etc.al., "Acoustic modeling for large vocabulary speech recognition",Computer Speech and Language 4, 1990,pp.127- 165.

[21] MatusPleva, StanislavOndas, JozefJuhar, Anton Cizmar, Jan Papaj and LubomirDobos, "Speech and Mobile Technologies for Cognitive Communication and Information Systems",2011.

[22] BeshoyMorkos, TorstenRilka, " Mobile devices within manufacturing environments: a BMW applicability

study ",Int J Interact Des (2012),10 April 2012 © Springer-Verlag2012,pp:101– 111.

[23] J.Ferguson, Ed., "Hidden Markov Models for Speech", IDA,Princeton, NJ,1980.

[24] B.H.Juang and S.Furui,"Automatic speech recognition and understanding: A first step toward natural human machine communication", Proc.IEEE, 88, 8, 2000, pp.1142-1165.

★★★