Artificial Speech Recognition

Alvee Ibne Rahman

CONTENTS

***Abstract*—In recent years, Artificial Speech Recognition Systems has been a popular topic in technology due to its use in voice assisting software in smart phones. But it is a new technology with a lot of development necessary to refine it. This paper provides an overview on Automatic Speech Recognition, how it works, its applications, the innovations made in this field by outlining the major themes and advances made in the past decade on this field. Speech accuracy remains an important research challenge. The existential problems and various techniques to solve them are presented in this paper. This paper aims to compare some well-known speech recognition systems and give us a general sense to where this technology stand at the present.**

I. INTRODUCTION

Automatic Speech Recognition (ASR) is the process of converting an acoustic speech signal to a sequence of words, by means of an algorithm implemented as a computer program. Research in speech processing and communication was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of signal processing. Speech recognition made it possible for humans to give commands using speech and make the machines understand what we are asking them to perform.

II. TYPES OF SPEECH RECOGNITION

There are two types of speech recognition: speaker-dependent and speaker independent. Speaker dependent system are commonly used for dictation software whereas speaker independent software are commonly found in telephone application. Speaker dependent software learns the unique characteristics of its user's voice. This means that when a new user starts using the software, he or she must first train the software by speaking to it, so that the computer can analyze how the person talks. Typically the computer would prompt the user to read out some assigned texts before they use the speech recognition software. Speaker independent software is designed to recognize anyone's voice without the necessity of training them first. The only practical use of this application is interactive voice response system. The only downside in this speaker-independent software is generally less accurate than speaker dependent software. To overcome the error rate, the speech recognition engine limits users to limited number of grammars they can use. [1]

III. ASPECTS OF SPEECH RECOGNITION

There are four major aspects of speech recognition based on their ability to recognize different types of utterances- isolated

words, connected words, continuous speech and spontaneous speech. Isolated word recognizers requires the speaker to dictate one word at a time with a small pause in between. Connected word recognizers does not require a significant large pause time in between singe utterance. Continuous speech system allows users to speak almost naturally while the computer determines the content. Continuous speech recognizers are the hardest to develop since they need special methods to determine utterance boundaries. Spontaneous speech are speech that are naturally sounding and not rehearsed. ASR with spontaneous speech recognition should be able to handle a variety of natural speech features such as word running together like ums and ahs and even a little stuttering.

IV. Applications of ASR

Some of the main domain where ASR have been heavily implemented are the Telecommunication Sector, Education Sector, Domestic Sector, Military Sector, Medical Sector, Artificial Intelligence Sector, Linguistic Sector etc. The telecommunication sector is using ASR to assist callers with telephone directory inquiry without operator assistance. Teaching foreign students how to pronounce vocabulary correctly and enabling handicapped student to input text verbally without the use of keyboard are two of the most significant use of ASR in the education sector. The Artificial Intelligence sector have been heavily depending on ASR with robotics. Some other uses of ASR are language translation, speech to text conversion (removing human error while typing), travel, banking, Avionics, Automobile portal, and also in voice command in natural language user interfaces like Siri, Google Now and Cortana.

V. ASR Timeline

The development of ASR is dated back to Alexander Graham Bell when he created a recording device for dictation. That led to the creation of machines that emulated human speech in the 1800s. In the 1900s, these machines underwent improvement, which led scientists to think about phonetic sound range and how to create machines that could emulate it. Scientist developed machines that could recognize vowels and digits, thus the modern understanding of ASR was born. While ASR developed and changed over the century, it is only the last 30 years or so that has been formalized, popularized and widely used by the public. In the 1980s, voice recognition has been used for automating services over the phone. We have the iPhones Siri program which can query the web for simple search like directions and listing.

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the frequency and amplitude of a digital signal of speech in compressed form, using the information of a linear predictive model. LPC was formulated in the late 1960 by Atal and Itakura. [8] LPC greatly simplified the estimation of the vocal tract response from speech waveforms. Later, by mid-1970, Itakura, Rabiner, Levinson and others, proposed the basic ideas of applying fundamental pattern recognition to speech recognition, based on LPC methods. During this same time period, Tom Martin and his company developed the first real ASR product called the VIP-100 System. VIP-100 System influenced the Advanced Research Project Agency of the U.S. Department of Defense to fund the Speech Understanding Research program. Carnegie Mellon University (one of the contractors of the ARPA program) created a system called Harpy [2] which could recognize speech using a vocabulary of 1,011 words. Harpy used graph search where the language was represented as connected networks derived from spoken word representation and word boundary rules, which determines the beginning and the end of a word. Along with the ARPA-initiated projects, IBM and AT&T Bell Laboratories were working on the applicability of ASR for commercial applications. IBM created a speaker-dependent system, focusing on the size of the recognition vocabulary and the language model structure. AT&T Bell Laboratory created a speaker-independent system that deal with the acoustic variability intrinsic in the speech signals coming from many different talkers, often with various regional accents. Their major approach was keyword spotting which was the primitive form of speech understanding. In between 1980 to 1990, speech recognition research changed their methodology from a more rigid template-based model approach to a more complex statistical modeling framework.

A stochastic process is a collection of random variables, representing the evolution of some system of random values over time. A doubly stochastic model is a model that observes random variable and models them in two stages. In one stage, the distribution of the observed outcome is represented in a fairly standard way using one or more parameters. At a second stage, some of these parameters are treated as being random variables themselves.

The Hidden Markov Model (HMM) was developed at the mid-1980s. HMM is a doubly stochastic process, which models the intrinsic variability of a speech signal and structures the spoken language in an integrated and consistent statistical modeling framework. In simple words, HMM is a statistical model which contains processes with hidden states. The state is not directly visible, but output, depending on the state, is visible. When users speak the same word, the acoustic signals are different even though the linguistic structure in terms of pronunciation, syntax and grammar remains the same. A probability measurement using the Markov chain is used to represent the linguistic structure and a set of probability distribution is used to account for the variability in the acoustic realization of sound in an utterance.

The Baum-Welch algorithm was implemented on a training set, to obtain the best set of parameters that defines the corresponding model or models. These models are equivalent to training and learning. HMM addresses the features of a probabilistic sequence of observation which may not be a static function, but a dynamic function that responds to the Markov chain. Even though this technique is powerful, it is limiting. There is a log-concave density constraint, which has later been replaced by elliptical symmetric density constraint. This improved the performance on a speaker independent task. In the mid-1980s, HMM was combined with the finite state network [15], which has been a major component in almost all modern speech recognition systems till this date.

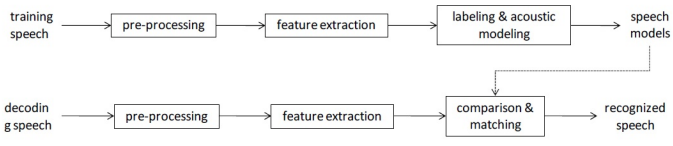
In the 1990s, progress had been made in the development of software tools that enabled individual research all over the world in ASR. The system was made by Cambridge University team and they called it the Hidden Markov Model Tool Kit (HTK). This system is one of the most widely adopted software for ASR till this date.

In the late 1990s, real speech enabled applications were developed. AT&T created an automated handling of operator-assisted caller called VRCP [8] which accepted tone touch input. As we draw closer to the 21st century, large vocabulary systems were built with unconstrain language models, and constrained task syntax model for continuous speech recognition and understanding. Also the fact that the vocabularies provided by the Web based profiling makes it likely that the user would use one of the 500 million entity entries stored by the Web search engine. The social graph used for web search engines greatly reduce the needed search space. Around 2000, a variational Bayesian (VB) estimation and clustering techniques were developed to solve the adaptive learning process for ASR. In 2005, improvement were made in Large Vocabulary Continuous Speech Recognition System which increased performance. In 2007, the difference in acoustic feature between spontaneous and read speech using large scale speech data base have been analyzed. Sadaoki Furui investigated SR methods that could adapt to speech variation using a large number of models trained based on clustering technique.

VI. Components of an Artificial Speech Recognition System

The process of speech recognition can be broken down into several phases. Sound patterns have to be recognized or classified into a category that represents a meaning to the user. This can be a problem computationally. Every acoustic signal can be divided into small basic sub-signals. Different levels of sound are created, where the top level contains complex sounds and the lower levels have more basic, short and simple sounds. The lowest level contains the most fundamental sound. An ASR system checks sound at the most basic level and uses probabilistic rules to determine what the sound represent. Once these sound samples are put back together into a complex sound on the upper level, a new set of deterministic rules is used to predict what the new complex sound represents. At the topmost level, the ASR system figures out the meaning of the complex expression. Automatic speech recognition (ASR) is performed by a computer algorithm designed to take a speech waveform as input and produce as output a useful transcription of that speech. Before recognition or decoding of the speech can be performed, the ASR system must be trained. All current speech recognition systems perform the same fundamental operations. These steps are depicted in Figure 1 for the training and the decoding phases of recognition.

Fig.1 illustrates the main processes in an ASR



For both training and decoding phases, the speech is converted in a form that can be utilized by the recognizer. Next, the desired features, those deemed essential to the recognition of speech are extracted.

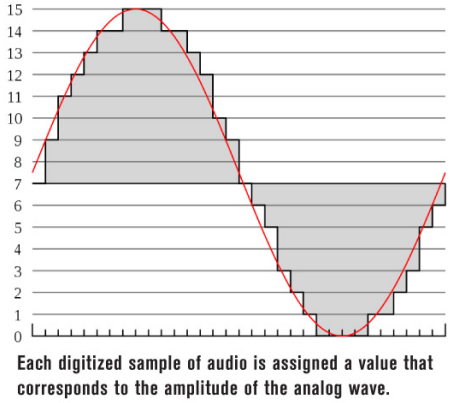
A. Training Phase and Decoding phase

During the training phase of recognition, these features are labeled, so that the region of the speech can be associated with one or more phonetic labels. An acoustic model is used to set up a relationship between similar parameters to reduce the number of independent parameters. In the decoding phase, the speech to be recognized is also pre-processed and the features are extracted. A comparison and a matching technique associates each region of the decoded speech with the training region deemed most similar.

B. Pre-processing

Spoken speech is causes vibration in the air. The Analog-to-Digital converter converts analog waves to digital data. The waves are sampled at about 8000 samples per second. The signals are segmented to create frames small enough to be stationary but large enough to contain enough information necessary for recognition. A filter is used to remove unwanted noise and stabilize the volume. Speech alignment is also carried out at this stage which fixes the speech speed.

Fig. 2 Shows an analog wave and its corresponding digital signal

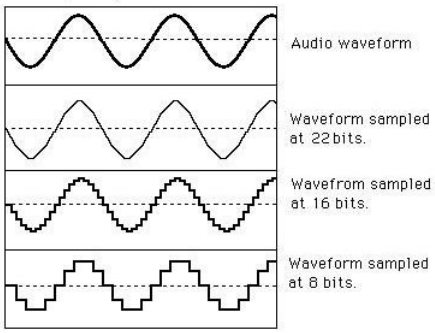


C. Feature Extraction

At this stage the speech signal is ready for the training or recognition process. The features deemed necessary for the recognition of speech are extracted from the signal. The characteristics of the desired feature should not vary greatly from one utterance to the next, nor from one speaker to the next for a particular phonetic unit. But it should change significantly from one phonetic unit to the next. These features include spectral characteristics of speech, which is the information about the energy, wavelength and amplitude of the in the signal speech. Energy is used to determine pitch, frequency and other parameters that may indicate the phonetic identity of the signal segment. Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates

of speech parameters. LPC is the principal tool for ASR and is used to convert a speech signal to a more suitable form for processing and evaluation. The Linear Prediction coder compresses a speech into a 2.4 kb/s.Quality of the speech is sacrificed, but this has to be done if a quick feedback is wanted. [6]

Fig. 3 shows sound quality and sample bit rate.



D. Labeling and acoustic modeling

Signal is divided into small segments as short as a few hundredths of a second. The program then matches these segments to known phonemes in the appropriate language. A phoneme is the smallest element of a language.

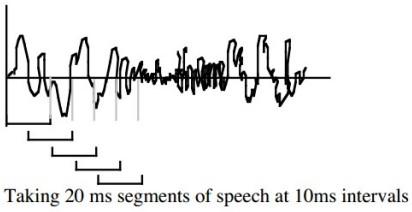
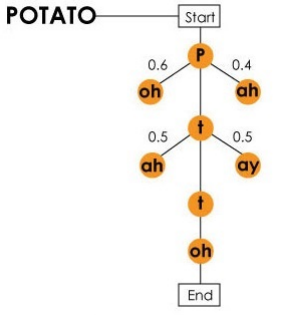


Fig. 4

For recognition purposes, speech is often modeled as the output of a Hidden Markov process in which the feature vectors of the decoding speech serve as observations emitted from a collection of states. This acoustic model, called a Hidden Markov Model (HMM), consists of the probabilities of transitioning from one state to another as well as the probability density functions of the observations of each state. Each phoneme has its own set of states and transition probabilities, which creates a reference model for that utterance. As the features are extracted, phonetic label is assigned. [5]

Fig. 5 A Markov Model For the Word ”POTATO”



E. Comparison and matching

Once the features are extracted, they must be compared to those of the training speech to determine which phoneme will produce an n-dimensional vector of numerical features (i.e. feature vector) that represent that particular phoneme. There are several comparison techniques. One comparison technique utilizes vector quantization in which the decoding vector is compared to the centroids of vector codebooks. The decoding vector is then represented by the codebook vector closest to it. Another common comparison method is template matching with dynamic time warping (DTW). This technique creates a best fit to each of the feature vector from the training data. A score is used to show the exactness of the match. The best score is considered to be the utterance actually spoken, provided that the score exceeds some predetermined threshold. The most common pattern matching technique that most mobile devices use now a days is called the Hidden Markov modeling. With a given utterance to be recognized, each feature vector of the decoding speech is scored using an algorithm which determines the model most likely to have been produced. The model with the highest probability is selected, provided the probability is above the given threshold.

VII. How ASR work?

Voice recording, acoustic signal processing, and recognition with the help of language models are the 3 main steps of ASR. The voice is being recorded by a sound receiver, typically a microphone. Then the sound is converted from analog signals to digital representation. This process is called digitization. Next, signal processing occurs where the speech is separated from background noise.

During recognition, an acoustic model is created by taking the audio recording and their text transcriptions and using software to create statistical representations of the sound that make up each word. This is used by the speech recognition engine to recognize speech.

A language model is also implemented mostly in speaker independent software. Language model is a file that contains the probabilities of sequences of words. This model is used for dictation applications, and voice response type application.

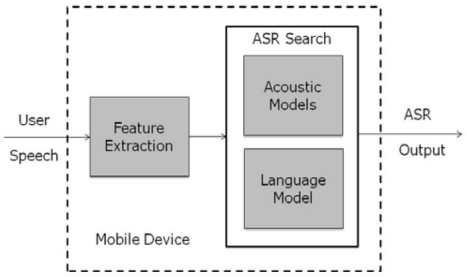
For any particular type of ASR, either the acoustic model or the language model or both might be used for speech recognition. Once the speech have been recognized, it’s the job of the speech engine to play back text in a spoken voice. The last step of ASR is the feedback which can either be displayed or spoken back, depending on what the user inquired. [5]

VIII. How Siri work?

Siri is an intelligent personal assistant and knowledge navigator which works as an application for Apple Inc.’s iOS. The application uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services. At first, the sound of the users speech is immediately encoded into a compact digital form that preserves the information in the speech. The ambient background noise from the signal is reduced and volume is optimized. It is crucial that the user have

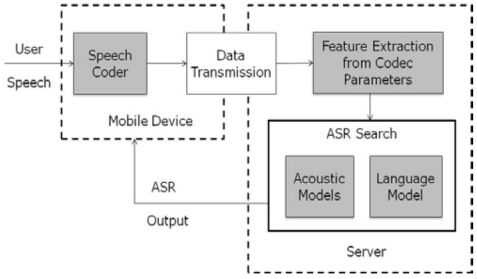
data connection on his phone. The digital signals are relayed wirelessly through a nearby cell tower and through a series of land lines back to your Internet Service Provider where it then communicated with a server in the cloud, loaded with a series of models honed to comprehend language. Simultaneously, the speech is being evaluated locally, on the device. A recognizer installed on your phone communicates with that server in the cloud to gauge whether the command can be best handled locally. Simple commands like asking Siri to play a particular song from the music library does not require an internet connection. If the local recognizer deems its model sufficient to process the users speech, it tells the server in the cloud that it is no longer needed. [9]

Fig. 6 ASR System for simple command



When you give a complicated command, the phone sends the bits of data to a central server, where it can access the appropriate software and corresponding database. The software analyzes the speech by breaks the speech down into tiny, recognizable parts called phonemes. The server compares the phonemes against a statistical model to estimate, based on the sounds the user spoke and the order in which the user spoke them, what letters might constitute it. At the same time, the local recognizer compares your speech to an abridged version of that statistical model. For both, the highest-probability estimates get accepted. [9]

Fig. 7 ASR System for complex command



Based on these opinions selected with the highest probability, the users speech is being understood as a series of vowels and consonants, which is then then run through a language model, which estimates the words that the users speech is comprised of. Given a sufficient level of confidence, the computer then creates a candidate list of interpretations for what the sequence of words in the speech might mean. Siri is mostly used for sending text messages hands free. When a user speaks a command, for example I am commanding Siri Message Mike Heroux I will be late for class today. Siri will go thought all the process mentioned above, and if there is enough confidence in the result, and there is –the computer determines that users intent is to send an text message, Mike Heroux is your addressee and therefore his contact information should be pulled from my phone's contact list and the rest is my actual note to him. If your speech is too ambiguous at any point during the process, the computers will defer to me, the user: Did you mean Mike Heroux, or Mike Hartz? Since we are focusing on the recognition aspect of Siri, we should not worry about how Siri distinguishes between a command and a comment. Siri will read out the text back to me and if everything looks okay, it will ask for my final confirmation for sending the text message. All these steps are completed in approximately 3 seconds.

IX. Major Challenges

A. Everyday audio

Current ASR systems returns a significantly degraded performance when they encounter audio signals which differ from the limited conditions under which they were originally developed and trained. We need to create and develop a system which would be more robust against variability and shifts in acoustic environment, reverberation, external noise sources etc.

B. Rapid Portability to Emerging Languages

The development of acoustic and language model in todays state-of-the-art ASR systems requires large collection of domain specific speech and text examples. For many languages, these sets of resources are not available, thus limiting the rapid development of ASR. We need a language-universal paradigm.

C. Self-Adaptive Language Capabilities

The way words are being pronounced, varies from person to person. Various statistical models that the ASR follows were made based on training data such as transcribed speech, and from human-supplied knowledge, such as pronunciation dictionaries. Sometimes these built-in knowledge becomes obsolete and that affects the performance of ASR. Retraining the ASR would be costly and extremely time consuming. So in order for the ASR to perform well consistently, researchers need to create a self-adaptive speech technology which will learn at all level of speech and language processing to cope with changing environments, non-speech sounds, pronunciations, dialects, accents etc.

X. Performance of ASR

Automatic speech recognition (ASR) systems are used in various applications in our day-to-day life. Rapid development in this field resulted in many systems and devices with voice input and output. Some example of ASR implementation are automated transcription of audio and video recordings, radio or TV inputs, devices in automobiles controlled by voice etc. Such wide application areas of ASR brings frequent use of such systems in noisy environment, so the issue of noise robustness have been an important topic in major research efforts. Performance of ASR is determined in terms of speed and accuracy. Accuracy is the percentage of words in a

command that the ASR accurately recognizes. Error rate is the percentage of words that the ASR fails to recognize correctly. The error rate increases as the vocabulary size in the ASR database grows. It has been seen that the numeric digits from 0 to 9 can be recognized almost perfectly. When the vocabulary size is 200, 5000, and 100000, the error rate may increase to 3%, 7% or 45% respectively. The English language contains alphabets which are difficult to discriminate because they are confusable, especially the E-set:B, C, D, E, G, P, T, V, and Z. An error rate of 8% is considered good for this vocabulary. [12]

The easiest way to measure accuracy is by calculating the word error rate (WER). Speed is measured with the real time factor. Other measures of accuracy includes Single Word Error Rate (SWER) and Command Success Rate (CSR). Word error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence. [13]

The Levenshtein distance is a string metric for measuring the difference between two sequences. The Levenshtein distance between two words is the minimum number of single-character edits required to change one word into the other. These single character edits could be insertion, deletion or substitution. [11] The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference word sequence using dynamic string alignment. Then we use this following equation

$$WER = 100 * (S + D + I)/N \tag{1}$$

where S is number of Substitutions, D is number of Deletions (omitted from the speech), I is number if Insertions and N is number of word in reference statement. The result is a numerical percentage of errors found in the statement.

Here is an example of how WER is being calculated:

Reference Statement: "portable **** PHONE UPSTAIRS last night so"

Statement displayed by the ASR: "portable FORM OF STORES last night so"

As we can see, there is one instance of Insertion, and 2 instances of substitution in the displayed statement. Also, the total number of word in reference statement is 6. With all these values calculated, the WER will be 100(1+2+0)/6 = 50A low WER means the ASR system has better performance and have higher accuracy at recognizing words. The opposite is true for ASR system having higher WER.

XI. Factors Affecting the Performance of ASR

A. Speaker Dependent ASR vs Speaker Independent ASR

The performance of speech recognition varies significantly depending on how well it copes with all the speaker dependent and speaker independent factors. ASR systems can be categorized by its training method into speaker-dependent and speaker independent systems. A speaker-dependent system is intended for use by a single user. The user can include new vocabulary into the system at the expense of training. The language and pronunciation behavior of the user are automatically taken into account. A speaker-dependent system has an error rate 2 to 3 times lower that speaker-independent system, due to individual training of the system, given that both systems have same amount of vocabulary and same amount of training time. A speaker-independent system is intended to use by any speaker. It does not require individual training. These systems comes with per-trained speech models on large amount of training data and they are predominantly Hidden Markov Model-based. Speech can be isolated, discontinuous or continuous. Single words are being used in isolated speech, making it easier to recognize by ASR systems. Full sentences separated by silence are used in discontinuous speech making it equally easier for ASR systems to recognize. It is much harder for ASR to recognize continuous speech since it consists of naturally spoken sentences. [1]

B. Speech Distortion

Distortion of speech occur when the speech is digitized (also known as speech coding). This can be countered through recognition of speech directly from the stream of bits, which helps eliminate the need for reconstructing of the speech signals and eliminating the distortion caused by it. There is a relative difference of 6.7% between the recognition error rate of un-coded speech and that of speech reconstructed from Mixed-excitation linear prediction encoded parameters (MELP). MELP is a United States Department of Defense speech coding standard used mainly in military applications and satellite communications, secure voice, and secure radio devices. The relative difference between the recognition error rate for un-coded speech and that of encoded speech recognized directly from the MELP bit stream is 3.5%. Distortions due to noise are offset through appropriate modification of an existing noise reduction technique called minimum mean square error log spectral amplitude enhancement. A relative difference of 28% exists between the recognition error rate of clean speech and that of speech with additive noise. Applying a speech enhancement front-end reduced this difference to 22.2%. [4] The spectral subtraction method is a simple and effective method of noise reduction. In this method, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved. It is assumed that the signal is distorted by a wide-band, stationary, additive noise, the noise estimate is the same during the analysis and the restoration and the phase is the same in the original and restored signal.

C. Background Noise

In noisy environments, speakers exhibit a reflexive response known as the Lombard Effect which results in targeted speech modifications. These modifications are not only due to increase in volume, but also due to changes in pronunciation which vary between speakers and based on the amount and type of noise. Studies suggest that the Lombard response is primarily automatic, and is involuntary. Speech variation caused by

ambient noise can be more degrading in terms of speech recognition accuracy than the noise itself. It is not possible to eliminate or selectively suppress the effect of Lombard speech. [10] Traditional algorithmic approaches to speech recognition need to be adapted to accommodate dynamic stylistic changes in speech signals that are brought about by mobile speech input under noisy conditions. The fundamental implication for speech recognition systems in the future is that they will have to be capable of handling variation in speech signals that come with mobile speech input. As stated previously, performance of ASR systems deteriorates quickly in the presence of background noise. ASR systems use a technique called spectral subtraction to deal with it. Spectral subtraction works best when the magnitude of noise spectrum is lower than that of the speech spectrum, which means that the speech should have a moderate to high signal-to-noise ratio. The results are not as favorable for speech of low signal-to-noise ratio. It has been observed that, in addition to spectral magnitude error, phase and cross-term errors exist in the spectral subtraction estimation. In signal processing, phase distortion is distortion that occurs when a filter's phase response is not linear over the frequency range of interest, that is, the phase shift introduced by a circuit or device is not directly proportional to frequency. Although the insignificance of phase in speech recognition accuracy is commonly accepted, results from human perception experiments have indicated that phase may significantly contribute to speech intelligibility. Furthermore, slight improvements have been made in the recognition of noisy speech through the incorporation of phase spectrum into the acoustic features. The impact of phase distortion on the accuracy of ASR systems is yet debatable, and therefore needs further examination.

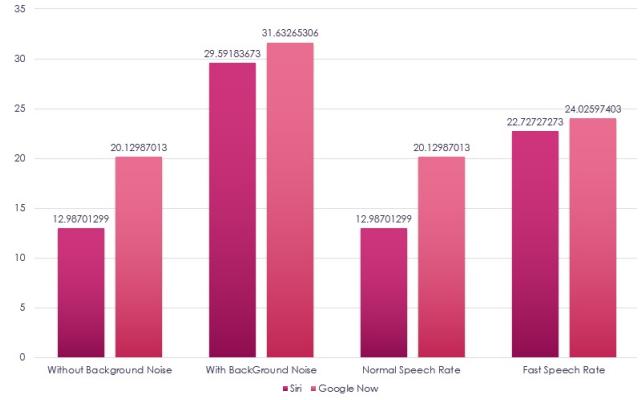
D. Speech Rate

The relationship between speaking rate variation and different acoustic correlates are usually not well taken into account in the modeling of speech rate variation for automatic speech recognition, where it is typical that the higher the speaking rate is, the higher the error rate is. Usually, slow speaking rate does not affect performance; however, when people hyper articulate, and make pauses among syllables, speech recognition performance can degrade a lot.

XII. Experimental demonstration of WER using ASR systems

In this experiment, I tested how background noise and rate of speech affect the accuracy of speech recognition.U have used two of the most popular ASR systems available in the market: Apples voice interface Siri and Googles voice search called Google Now. The dependent variables in this experiment is the word error rate, and the independent variables are the two ASR systems. I had 12 participants in my experiment, each providing me with 9 speech inputs. Firstly, I took speech samples from a participant who was instructed to read a short statement that I providing them. This experiment was conducted in a noise free environment. For the first try, I asked the participant to read the sentence normally. On the second try, I ask the participant to read out the statement at a much faster rate. On both cases, I recorded the amount of time it took for them to read out the statement. I calculate the WER for both cases using the equation that I previously provided. I have repeated the same experiment, except this time I conducting it in a public place having background noise. I predicted that the WER will be lower when the speech is spoken in a setting where there is no background noise and speech speed is normal. I also expected to see that WER will be highest where the speech speed is high along with background noise. With all the data I have collected, I was able to create a graph showing how speech rate and background noise directly affect the performance of ASR.

Fig. 8 Graph showing WER at different conditions



I compare the WER of both Siri and Google Now under all the mentioned conditions. Siri showed a WER of 12.98 at normal setting with no background noise at normal speech rate whereas Google showed a 20.12. As we introduce background noise, the WER jumps upto 29.6 for Siri and 31.6 for Google. When speech rate increased, the WER also increased to 22.7 for Siri and 24.0 for Google With these data, we can say that Siri has lower WER compared to Google, which means it has more successful recognition that Google, thus it performs better than Google.

XIII. Future Works in ASR

Even though ASR made huge progress in the last 30 years, two main obstacles persists; space complexity of the algorithms and speaker variability. Error rates for databases of conversational speech remains unacceptably high. For practical purpose, errors must be reduced further before ASR gets heavily implemented in everyday life. An ASR containing a very large lexicon would be impractical to search, a large amount of memory is needed to store the speech, phonemes and words. Speaker variability is a bigger issue, because errors are made when interpreting speech due to the fact that, no two speakers sound exactly the same, and that even a single speaker never say the same word the same way twice. ASR systems must be very robust to handle such variability. Speech recognition process may work well in clean conditions but it degrades significantly in speaker and channel mismatch conditions. Speech enhancement method needs to be improved, which can carry out spectral subtraction, improves

performance in noisy conditions. We should expect speech recognition in the next 40 years passing the Turing Test [7] bridging the gap between us and machines.

XIV. Present Status of ASR

ASR rapidly gained popularity as soon as they were introduced in handhold devices. Google, Apple and Microsoft are the major companies manufacturing smart phones which have in-built ASR systems. Even though the popularity of ASR is growing, the practical implementation of ASR have not grown proportionally due to performance issues that ASR faces. The recognition rate is not high enough to replace traditional user interfaces like the keyboard and touch screen command. In order to make ASR more useful, developers have been trying out different recognition technique to improve the recognition rate and improving algorithms which helps the system understand speech in noisy environment.

XV. Future Trends

A. Recognition of Speech, Body Language, Facial Expression

Robotic android projects are being carried out by the US and Japan where facial expression are being used to create an emotional bond between humans and machines. Speech recognition systems are taught not only to recognize speech but also read body language and facial expression. This information can be used for threat assessment. This technology will be useful in high security places like airport and border crossing, replacing human workers at those locations or check point. If a speaker smiles at a robot android and it smiles back at him while having a conversation, this ups the emotional value of the conversation to the user. Perhaps, the system might complement the user. In the future, ASR systems will have their individual personality, they can mirror responses or reciprocate an angry response or work to diffuse a situation, all depending on how they are programmed. Within five years, we can expect the software to have the ability to sense emotion, hesitation, aggression, hostility, anger, etc. Haptic is another field of science, which is a form of interaction that involves recognition through touch. In the future, robots being built might look like humans and mimic human characteristics. With the help of Haptic and Voice and Facial Recognition, the robots will have better performance in daily life interactions with humans. For example, if a robot feels a strong handshake along with a self-confident voice of an individual, the robot might elevate the trust factor about the person and interpret the person as being authoritative. In the future, we would expect most interaction with robots is through gestures and two-way natural-language spoken communication.

B. Emotion and Humor

In recent studies, scientists are trying to emulate emotion and empathy. This feature of speech could be useful in automated call center. Now a days, call centers have ASR which navigate us just through the menu section, but as Artificial Intelligence of machine increases, they will be able to deal with customer's queries or concerns and might be able to respond appropriately. ASR could also be implemented for crisis hot lines in the future. Present day ASR systems do not understand humor. As Artificial Intelligence improves, software engineers will create joke recognition systems, where the computers will understand irony and know when a user is telling a joke, then reciprocate with a joke of their own, perhaps creating a joke from scratch. The system would be per-loaded with all the jokes common to human interaction in all cultures. It will be able to pick one that has most likely not been heard by the user they are working with at the time. The system should also keep track of jokes that have already been told and make sure it does not repeat it. This means advances for human companionship for long-term space flight, help with rehabilitation and ease off tension of humans working alongside robots.

C. Military Implementation

Advanced research in the US Military is working on a new technique where instead of speech input, they are developing a technique which allows vocal cords to be recorded without actually speaking. This is done by implanting a device near the larynx which reads the sensitive vibrations. Transmitters are used to send the signals. A tiny ear piece reads out the speech to another user. This systems will develop and soon the secret service members, Special Forces, SWAT teams will be able to communicate without being heard by civilians. [14]

D. Other Future Trends

As we have seen in the last 30 years, the performance of ASR have improved steadily, so it will not be surprising to see greater improvement in performance rate in the future. The simplest way for current state-of-the-art recognition systems to improve performance on a given task is to increase the amount of high quality task-relevant training data from which its models are constructed. Due to the Internet, there are a lot of open source data that the ASR could use for training. Maybe not in the next 5 years, but there will be a point when the ASR can self-train itself and will have enough data to have equal recognition rate as a human. Once this is achieved, we would expect people from different part of the world talking to each other over translating telephones, even if they do not speak the same language. In the aviation sector, automatic airline reservation by voice over the telephone will be a norm. Speech recognition will be commonly used at home for voice interactive television, control of home appliance, and home management system. It has already been seen that some hybrid cars are using ASR for simple command like playing music or radio. In the near future, ASR will also be able to carry out command like opening window, lock doors, or turning on headlights. Also, most speech recognition systems will abandon the Hidden Markov Model and n-gram paradigm and move over to acoustic and language modeling. Also the majority of the text created will be through continuous speech recognition, instead of isolated or discrete speech recognition.[3] Accuracy has been one of the biggest challenge in ASR. There has not been a single system which is 100% accurate at speech recognition. In the future, the accuracy rate

will become better due to better error detection and better error fixing after detection. Dictation speech will gradually become accepted. ASR will be used in data entry. Natural user interfaces will be better at providing us with better answers to our queries. With the improvement in artificial intelligence, ASR can make greater use of it to attempt to guess what the speaker intended to say, instead of what was actually said. Technological advancement in microphone and sound system will help adapt more quickly to changing background noise levels, different environments, with better recognition of extraneous material to be discarded.

XVI. Conclusion

In this paper, I have talked about what ASRs are and the types of ASR present at the moment. I have distinguished the two main types of ASRs, speaker dependent and speaker independent systems. Next, I talked how researchers, and government funded researches led to the progress of creating ASRs. In order to understand how ASRs function, I broke down the ASR system and explained each individual components and how they contribute in recognizing speech. In this case, I have used Apple's Siri as an example. Accuracy of ASR have been an issue since its inception. I have discussed the major challenges, that the researches are still facing, while dealing with performance. I have also explained how performance of ASR is measured and all the factors that affect it. My experimental demonstration determined that having background noise and higher speech rate reduces accuracy of ASR systems. In the future we would expect ASR systems to have better recognition and would also be a common interaction method between humans and machines.

References

[1] M. A. Anusuya and S. K. Katti. Speech recognition by machine, A review. *CoRR*, abs/1001.2267, 2010.

[2] M.A. Anusuya and S.K. Katti. Speech recognition by machine: A review. *IJCSIS International Journal of Computer Science and Information Security*, 6-3, 2009.

[3] Janet Baker, Li Deng, Sanjeev Khudanpur, C Lee, James Glass, and Nelson Morgan. Historical development and future directions in speech recognition and understanding. *MINDS Report of the Speech Understanding Working Group, NIST*, 2007, 2006.

[4] Jayne Angela Beauford. *Improving the automatic recognition of distorted speech*. PhD thesis, University of Pittsburgh, 2010.

[5] Ed Grabianowski. How speech recognition works. http://electronics.howstuffworks.com/gadgets/high-tech-gadgets/speech-recognition.htm.

[6] Serajul Haque, Roberto Togneri, and Anthony Zaknich. Perceptual features for automatic speech recognition in noisy environments. *Speech communication*, 51(1):58–75, 2009.

[7] Xuedong Huang, James Baker, and Raj Reddy. A historical perspective of speech recognition. *Commun. ACM*, 57(1):94–103, January 2014.

[8] B.H Juang and Lawrence R. Rabiner. Automatic speech recognition - a brief history of the technology development. *IEEEElsevier Encyclopedia of Language and Languistics*, 2005.

[9] Anuj Kumar, Anuj Tewari, Seth Horrigan, Matthew Kam, Florian Metze, and John Canny. Rethinking speech recognition on mobile devices.

[10] Joanna Lumsden, Irina Kondratova, and Scott Durling. Investigating microphone efficacy for facilitation of mobile speech-based data entry. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, pages 89–97. British Computer Society, 2007.

[11] Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. Predicting human perceived accuracy of asr systems. In *INTERSPEECH*, pages 1945–1948, 2011.

[12] David S Pallett. Performance assessment of automatic speech recognizers. *J. Res. Natl. Bureau of Standards*, 90:371–387, 1985.

[13] Josef Rajnoha and Petr Pollák. Asr system in noisy environment: Analysis and solution for increasing noise robustness. *Radioengineering*, 20(1):74–84, 2011.

[14] Douglas A. Reynolds. Automatic speaker recognition: Current approaches and future trends.

[15] R. Solera-Ureña, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-De-María. Svms for automatic speech recognition: A survey. pages 190–216, 2007.

Appendix

While working on this thesis paper, I have learned how to find scholarly resources from proper sources and properly cite them. I have learned how to narrow down the thesis. I have familiarized myself with TeX studio and will be using this useful tool in the future.I have used Microsoft Excel to aid me during data collection during my experimental demonstration. I found it particularly helpful when Professor Heroux gave us exercise for making better Abstract and Introduction. Also, his feedback on the presentation dry-run have been constructive. My overall experience while completing this thesis have been great! Although I have struggled at times, but I think I have learned a lot. Hopefully, in the future I will be able to do presentations on a professional manner and write better thesis papers. I have learned a lot about ASRs and I plan on keeping up-to-date with this field of science.