PERFORMANCE EVALUATION OF REAL-TIME NOISY SPEECH RECOGNITION FOR MOBILE DEVICES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
MIDDLE EAST TECHNICAL UNIVERSITY
BY

YASER YURTCAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

FEBRUARY 2019

Approval of the thesis:

PERFORMANCE EVALUATION OF REAL-TIME NOISY SPEECH RECOGNITION FOR MOBILE DEVICES

Submitted by **YASER YURTCAN** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin _____
Dean, **Graduate School of Informatics**

Prof. Dr. Yasemin Yardımcı Çetin _____
Head of Department, **Information Systems**

Assoc. Prof. Dr. Banu Günel Kılıç _____
Supervisor, **Information Systems, METU**

Examining Committee Members:

Assoc. Prof. Dr. Altan Koçyiğit
Information Systems Dept., METU _____

Assoc. Prof. Dr. Aysu Betin Can
Information Systems Dept., METU _____

Assoc. Prof. Dr. Banu Günel Kılıç
Information Systems Dept., METU _____

Assoc. Prof. Dr. Pekin Erhan Eren
Information Systems Dept., METU _____

Assist. Prof. Dr. Mustafa Sert
Department of Computer Engineering, Başkent University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: **YASER YURTCAN**

Signature :

ABSTRACT

PERFORMANCE EVALUATION OF REAL-TIME NOISY SPEECH RECOGNITION FOR MOBILE DEVICES

Yurtcan, Yaser

M.S., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Banu Günel Kılıç

February 2019, 67 pages

Communication is important for people. There are many available communication methods. One of the most effective methods is through the use of speech. People can comfortably express their feelings and thoughts by using speech. However, some people may have a hearing problem. Furthermore, understanding spoken words in a noisy environment could be a challenge even for healthy people. Speech recognition systems enable real-time speech to text conversion. They mainly involve capturing of the sound waves and converting them into meaningful texts.

The use of speech recognition on mobile devices has been possible with the development of cloud systems. However, delivering a robust and low error rate speech recognition system in a noisy environment still is a major problem. In this study, different speech samples have been recorded using a compact microphone array in noisy environments and a data set has been created by processing them through a real-time noise cancellation algorithm. A portable design of a mobile system with noise cancellation hardware and software was proposed to convert spoken words to a meaningful text.

Comprehensive tests were performed on several clean, noisy and denoised speech samples to measure the speech recognition performance of different cloud systems, noise robustness of the proposed system, the effect of gender on the speech recognition performance, and the performance improvement. The experimental results show that the proposed system provides good performance even in a noisy environment. It is also inferred from the results that in order to apply speech recognition using cloud based

systems on mobile devices, the noise level has to be low or real-time noise cancellation algorithms are needed. The proposed system improves speech recognition accuracy in noisy environments. Thus, the achieved performance and portable design together enable the system to be used in daily life.

Keywords: Speech Recognition, Speech Processing, Cloud Systems, Word Error Rate, Mobil Devices

ÖZ

MOBİL CİHAZLARDA GERÇEK ZAMANLI GÜRÜLTÜLÜ KONUŞMA TANIMA PERFORMANS DEĞERLENDİRİLMESİ

Yurtcan, Yaser

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Doç. Dr. Banu Günel Kılıç

Şubat 2019 , 67 sayfa

İletişim insanlar için önemlidir. Birçok iletişim kurma yöntemi bulunmaktadır. Bunlar arasında en etkili olanı konuşmadır. Konuşma ile insanlar duygularını ve düşüncelerini rahat bir biçimde ifade edebilmektedir. Bununla birlikte, bazı insanların işitme problemi olabilir. Dahası, gürültülü bir ortamda konuşulan kelimeleri anlamak sağlıklı insanlar için bile zor olabilir. Konuşma tanıma sistemleri, metin dönüşümüne gerçek zamanlı konuşma sağlar. Konuşma tanıma sistemleri genellikle ses dalgalarının yakalanmasını ve anlamlı metinlere dönüştürülmesini içerir.

Mobil cihazlarda konuşma tanıma kullanımı, bulut sistemlerinin geliştirilmesi ile mümkün olmuştur. Ancak, gürültülü ortamlarda gürbüz ve düşük hata oranlı konuşma tanıma sistemi sağlamak hala önemli bir sorundur. Bu çalışmada, gürültülü ortamlarda kompakt bir mikrofon dizisi kullanılarak farklı konuşma örnekleri kaydedilmiş ve gerçek zamanlı bir gürültü engelleme algoritmasıyla işlenerek bir veri kümesi oluşturulmuştur. Konuşulanları anlamlı bir metne dönüştürmek için gürültü engelleme donanımı ve yazılımı olan taşınabilir bir mobil sistem önerilmiştir.

Farklı bulut sistemlerinin konuşma tanıma performansını, önerilen sistemin gürültüye dayanlıklılığını, konuşmacının cinsiyetinin konuşma tanıma performansına etkisini ve performans iyileştirmeyi ölçmek için temiz, gürültülü ve gürültüden temizlenmiş konuşma örnekleri üzerinde kapsamlı testler yapılmıştır. Deney sonuçları, önerilen sistemin gürültülü ortamlarda bile iyi performans sergilediğini göstermektedir. Sonuçlardan ayrıca anlaşılmıştır ki, mobil cihazlarda bulut tabanlı sistemleri kullanarak konuşma tanıma yapmak için gürültü seviyesi düşük olmalıdır veya gerçek zamanlı

gürültü iptali algoritmalarına ihtiyaç duyulmaktadır. Önerilen sistem gürültülü ortamlarda konuşma tanıma doğruluğunu arttırmaktadır. Böylece, elde edilen performans ve taşınabilir tasarım, sistemin günlük hayatta kullanılmasına olanak sağlamaktadır.

Anahtar Kelimeler: Konuşma Tanıma, Konuşma İşleme, Bulut Sistemler, Kelime Hata Oranı, Mobil Cihazlar

To My Family

ACKNOWLEDGMENTS

I would like to thank my supervisor Associate Professor Banu Günel Kılıç for her support and guidance in this long and exhausting work. This study has also changed the way I look at the academic world. I cannot forget to thank my colleagues from ASELSAN for their technical support. Lastly, I want to thank my family who always provided motivation and morale during this time. This thesis is devoted to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANC	Active Noise Cancelling
AES	Advanced Encryption Standard
API	Application Programming Interface
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CNTK	Computational Network Toolkit
CWR	Correct Word Rate
DARPA	Defense Advanced Research Agent
dB	Decibel
DNN	Deep Neural Networks
DSP	Digital Signal Processor
EM	Expectation Maximization
GMM	Gaussian Mixture Model
GRPC	Google Remote Procedure Call
HMM	Hidden Markov Model
Hz	Hertz
ICA	Independent Component Analysis
LPA	Linear Predictive Analysis
LPC	Linear Predictive Coefficients
MEL	Mel-Scale Cepstral Coefficient
MEMS	Micro-electro Mechanical Systems
MFCC	Mel-Frequency Cepstral Coefficient
NCA	Noise Cancellation Algorithm
OS	Operating System
PLP	Perceptual Linear Predictive
PSCR	Public Safety Communications Research Group
REST	Representational State Transfer
SD	Speaker Dependent
SER	Sentence Error Rate
SNR	Signal To Noise Ratio
SI	Speaker Independent
SIRI	Speech Interpretation and Recognition Interface

SUR	Speech Understanding Research
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
ULA	Uniform Linear Arrays
WAR	Word Accuracy Rate
WER	Word Error Rate
WiFi	Wireless Fidelity

CHAPTER 1

INTRODUCTION

Communication is vital for human beings. In today's world, there are many ways to communicate information. Generally, there are three types of communication: Oral/speech, written, and body language. Speech is the most efficient form of communication that enables humans to share their thoughts and ideas. It is also a fast communication type that leads to instant feedback. Humans would not be able to describe many feelings without speech. However, it is sometimes difficult to understand what is spoken, especially in a noisy environment. Furthermore, this problem could be a challenge leading to many negative effects. Speech recognition systems overcome this problem by enabling real-time speech to text conversion.

Speech recognition systems enable people to understand spoken words to a certain level in noisy environments. General uses of these systems are voice dialing, command and control, dictations, and aided communication and monitoring. In the past five decades, speech recognition technology has made significant progress. Initially, the systems were not sufficient to provide robust solutions with a low error rate. Improvement in processors' computing power, development of advanced algorithms, the invention of better noise performance microphones, and availability of a large speech text data set contributed to this progress. These contributions enabled researchers to develop complex systems to analyze sounds and ensure correct word recognition. Modern speech recognition systems involve many subsystems. They include microphones to capture sound waves, and cloud computing systems to convert sounds to basic language units and construct words from phonemes. Over the past four decades, researchers have attempted to develop robust systems with low error rate. The key indicators of successful speech recognition systems are the low error rate, robustness and real-time operation.

Today's solutions make it possible to use speech recognition systems in our daily lives by utilizing mobile devices. Apple's Siri (Speech Interpretation and Recognition Interface) and Samsung's Bixby are the best examples of mobile device applications. People can use these systems to find out where the nearest restaurant is, to set alarms, to call people, to read emails, and much more. These systems are designed to work on a command and control basis. For example, the user gives a command and waits for an action. In addition to these systems, there are applications that translate spoken words instantaneously into text. The major technology companies Google, Microsoft, and IBM have such applications and these applications work with cloud systems. These applications instantly translate given speech into text and do not take

any action like Apple's Siri and Samsung's Bixby. In addition, such systems work in an unlimited dictionary compared to other systems.

Speech recognition systems on mobile devices generally provide sufficient results in a quiet environment but provide insufficient results under noisy conditions. In a noisy environment, the problem of speech recognition with a low error rate still persists. In this study, we have developed a system to overcome the speech recognition problem on mobile devices in noisy environments. This system allows real-time speech recognition with a low error rate up to a certain noise level.

In this chapter, problem definition, motivation, the scope of the thesis, the structure of the thesis and the objectives of the thesis are presented.

1.1 Problem Definition

As stated earlier, communication is crucial for human beings. Unfortunately, many people lose the ability to understand spoken words in noisy environments, especially, elderly people and people that have a hearing problem. Even healthy people can have difficulties in understanding speech in environments with a noise level above 80 decibels (dB). Any unwanted audible sound is called noise. In communication, the noise level is measured by signal to noise ratio and expressed as S/N or SNR. This ratio is measured in dB and is found by the following formula SNR;

$$SNR = 10log\frac{Ps}{Pn} \tag{1.1}$$

where Ps is the power of the signal and Pn is the power of the noise. If Ps and Pn are equal, the SNR is equal to 0 and the noise level is competing with the signal. So, what is the meaning of noise? Although there is more than one description of the noise, it is basically referred to as any unwanted disturbing sounds.

The noise is context dependent. For example, if two people are speaking simultaneously which one is noise depends on the context and the listener. The main problem in such situations is the presence of background noise and more than one speaker. There are different types of noise such as mechanical noise, traffic noise, people noise, and loud music, etc., which people are exposed to in their daily lives. Noise makes it more difficult to have a conversion and thus people need to give more attention to the speaker, which causes listener fatigue. The effect of background noise on speech recognition is more detrimental for older people [1].

Noise level is also an important problem for automatic speech recognition systems. The SNR is the main factor that affects the speech recognition performance [2]. The higher the SNR, the higher the quality of the incoming signal. Since the environment where mobile devices are used cannot be controlled, background noise is a major problem for speech recognition on mobile devices.

Most speech recognition applications on mobile devices are context dependent which means they try to perceive the speech as a meaningful sentence. For example, if the recognized sentence is "What is the weather life", it is changed to a meaningful form as "What is the weather like?".

Today's speech recognition systems provide sufficient results in quiet environments, but in noisy environments, the results are more than 100%. It would be nice to have speech recognition applications that show the same performance in both noise-free and noisy environments.

In our study, we aim to improve the performance of speech recognition on mobile devices in noisy environments. For this purpose, a compact microphone array is used for source separation to remove unwanted noise before speech recognition. We have also developed an application as a hearing assistant which shows what is spoken on the screen in real-time. Results show that the overall system is superior to standard ones.

1.2 Motivation

The main motivation of this study is to overcome speech recognition problems in noisy environments which have been worked on for the past the 50 years. The study aims to develop a portable mobile system that increases speech intelligibility and provides a better speech recognition rate. By using the proposed system and its portable feature, we want to overcome the problem of speech recognition in any noisy environments up to a certain level. Thus, people with hearing problems can gain the ability to understand what is spoken in the environment with the help of our designed system.

1.3 Objectives of the Thesis

This study has the following objectives:

- To find out which cloud system provides better speech recognition performance.
- To measure the effect of the noise level on speech recognition performance.
- To examine how robust the designed system is to noise.
- To investigate the effect of speaker gender on recognition performance.
- To quantify the performance improvement achieved with the developed noise cancellation algorithms.

1.4 Scope of the Thesis

The aim of this study is to improve noisy speech recognition performance on mobile devices. This study approaches the problem as a system design issue and integrates suitable hardware and software components to achieve the desired results. Therefore, improving the existing noise cancellation or speech recognition algorithms is beyond the scope of this thesis.

1.5 Structure of the Thesis

The rest of this thesis is organized as follows:

In Chapter 2, we provide an overview of the speech recognition technology, explain how speech recognition relates to deep learning, explain and compare cloud systems' performance, describe challenges, explain conventional noise cancellation methodologies, and present metrics for evaluating speech recognition performance.

In Chapter 3, the proposed system is described in detail, explain the specifications of the noise cancellation algorithm used, state reasons of selected transfer media and describe an application of speech recognition.

In Chapter 4, experimental setup is explained together with, covered speech recognition factors.

In Chapter 5, detailed results are provided and discussed.

In Chapter 6, concludes the thesis.

CHAPTER 2

LITERATURE REVIEW

In this chapter, an overview of speech recognition, deep learning for speech recognition, relation with cloud-computing, speech recognition on mobile devices, challenges for applications using speech recognition, noise cancellation methodology, audio transmission to a mobile device, and evaluation of speech recognition performance are investigated.

2.1 Overview of Speech Recognition

2.1.1 What is speech recognition?

As the name indicates, speech recognition is translation of spoken words into text. A speech recognition system basically captures sound signals, makes some process on them and converts them into text. The term "speech recognition" has been used since the early 1950s, when Audrey and his team at Bell Labs designed a machine capable of understanding spoken digits [3]. The machine had limited accuracy that was speaker-dependent. Since that time, there have been many breakthroughs in technology. In the early 1950s, computers had limited computational power and limited training data. Machine learning had not been introduced; there were no advanced algorithms; and no high-tech microphones were present. Now there are available powerful computers that perform millions of operations per second, high-tech microphones such as microelectromechanical systems (MEMS) microphones, cloud-computing technology, and improved learning techniques, including deep learning.

Adopting technological improvements has led to higher performance achievements that deliver robust and low error rate speech recognition systems. Apples SIRI (Speech Interpretation and Recognition Interface), Microsoft's Cortona and Google's Voice Search are prominent examples. These are very popular applications that enable users to interact with mobile devices via voice command. They are also internally linked with web search engines (Google and Microsoft Bing) that indexed the entire web [3] which allow the users to search for such things as the nearest restaurants, today's weather, and other information. Speech recognition has evolved from understanding spoken digits to understanding the meaning of what is said and facilitating the taking of appropriate action.

The basic speech recognition system consists of three main components, as illustrated

in Figure 2.1.

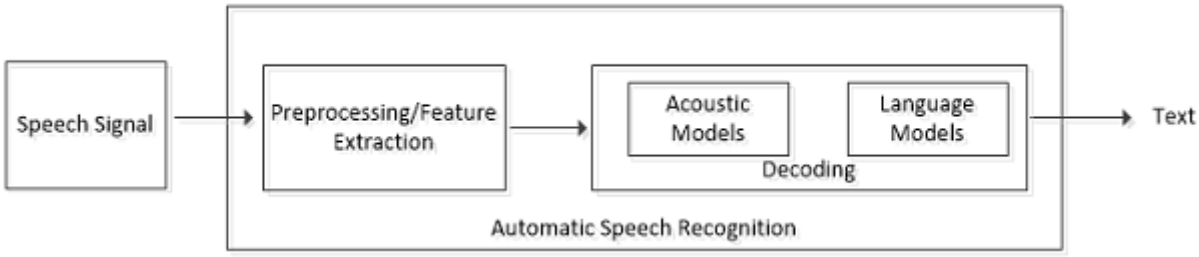


Figure 2.1: The Components of a Basic Speech Recognition System

The components of basic speech recognition systems are introduced in the next section.

2.1.1.1 Preprocessing and Feature Extraction

Preprocessing is the first step in speech recognition systems. In the following steps, digital format of speech signals are needed. However, the captured (recorded) speech signals are analog and they need to be transformed into a digital format for further analysis and processing. Transferring analog signals into digital format and applying basic filtering technique to remove some artefacts are called *preprocessing*. Feature extraction is the most important part of speech recognition. Good feature extraction can increase speech recognition performance. Feature extraction reduces the variability of speech signals since the speech signals have the changing characteristic over time [4]. It extracts the required significant parameters of speech signals and eliminates irrelevant unimportant parameters/features while dividing the speech signals into short frames (generally 20-25 ms duration and shifted 10 ms) [5]. By doing so, a quantitative representation of the speech signal is achieved for further processing. An important point is that the frames must be short duration so that speech signals can be viewed as stationary. Some of the extracted parameters are information on the speaker and the recognition of utterances. There are many features, such as Mel-frequency cepstral coefficients (MFCC) [6], Mel-scale cepstral coefficients (MEL) [7], Linear Predictive Coefficients (LPC) [8] obtained Linear Predictive Analysis (LPA) [4], and Perceptual Linear Predictive Coefficients (PLP) [4].

MFCCs are the most popular technique. They provide high accuracy with low complexity [6]. They are based on the variations of human hearing. Their performance is more sensitive to background noise and the number of filters used [9]. MEL models approximately the human hearing by scaled frequency. The frequency either scaled linear or algorithmic [7].

LPC is a method which provides robust and high accuracy of speech features efficiently by reducing required information on speech signal [8]. LPA is a static feature extraction method that is based on the assumption of past speech samples. The idea is that the current speech sample can be described by observation of past samples over a duration. However, it can not clearly recognize the words with similar utterances, because of the inherent assumptions. Different bit rates, the delay of the system, and

computational complexity affect the performance of the LPA [4].

PLP eliminates artefacts and hence improves speech recognition performance. It is short duration There are mainly three aspects: the critical-band resolution curves, the equal-loudness curve, and the intensity-loudness power law [7]. There are some common part with LPC. However, PLP is more efficient.

2.1.1.2 Decoding and Text

Decoding is the process of recognizing the text equivalent of the speech by using the output of the feature extraction. There are two types of decoding models, acoustic models and language models.

An acoustic model is the main part of the speech recognition system and is also called as the pronunciation model. These models provide a statistical representation of the sounds that make up words [10]. They play a very critical role in achieving a noise robust and high accuracy recognition system. They provide a relation between a speech sound and its corresponding phonetics. Thus they need to be trained with very large datasets that include various speakers of various ages and genders to provide a robust speaker independent system. There are several available acoustic models. Most widely used ones are the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM).

HMM is a widely accepted and feasible acoustic model used since the 1980s [11]. It is a statistical model that divides the obtained feature vectors into states. The states represent phoneme units of each word. For instance, the word "when" consists of "wh", "e", and "n" phone units. Each phoneme unit has different features with different distribution that is directly affected by the previous and next state. So, each phoneme HMM consists of three states and the "when" HMM has 9 states. Thus HMM has a set of different states that represent the characteristic of sound signal in order to find the relation of one state to another to make up the corresponding the word. HMM needs to be trained with a large amount of acoustic data to find the correct phone units. A large acoustic data set for HMM significantly reduces the recognition time.

GMM is a statistical model. Gaussian distribution are evaluating mean, variance and weight for representing GMM [12]. GMM is estimated as the probability density function. It is computationally efficient and easy to be implemented. It considers sound signals as consists of the sum of several independent components. GMM determines the relationship between input and states of HMM by means of expectation-maximization (EM) [12].

Language models calculate the probability of next sequence of words [13]. The aim is to determine the most suitable sequences of words from the signal. It is a statistical model, because the assumption of the next sequence is required by utilizing a training data set. The accuracy of the correct assumption is closely related to the training data set. Language models are language-specific and each language has its own limitations and characteristics.

The most commonly used language model in speech recognition is *the n-gram language* model. There are available other language models including bi-gram and tri-gram. Language models in speech recognition systems help to predict the best next-word sequences by considering previous n-1 words. It is thus used to distinguish similar word groups.

Language models decide on the next possible word considering the previous word and the training data set. The previous word is crucial, because it provides information on what the next word should be to follow the previous one. For example, if we examine the following sentence, "What is weather ...?", what should be the last word in the sentence? (like or life). In this case, the used language model and the data set play an important role.

In the bi-gram model, the probability of the next word depends on only the previous single word. So, the probability equations of the next word should be:

$$P(weather|life) \tag{2.1}$$

$$P(weather|like) \tag{2.2}$$

In the tri-gram model, the probability of the next word depends on the previous two words. So, the probability equations of the next word should be:

$$P(is, weather|life) \tag{2.3}$$

$$P(is, weather|like) \tag{2.4}$$

In the n-gram model, the probability of the next word depends on only the previous n-1 words. The choice of n depends on an application and number of words in the sentence. It is more suitable for long sentences. Generally, the previous three or four words provide the necessary information.

2.1.2 History of Speech Recognition

The first speech recognition system, namely the Audrey machine, was invented at Bell Laboratories in 1952 [14]. Some of its features [14];

- It was a fully analog system.
- It could understand only words of digits with pauses in between.
- It was a speaker dependent system and recognized digits spoken by a single voice who already adjusted to the system.
- Achieved 97-99% accuracy with the dependent speaker.

From the 1950s to the 1960s, limited digits and numbers could be recognized with speaker-dependent systems [15].

The 1970s decade saw many innovations in the speech recognition area. Continuous speech recognition was introduced, where the user was not required to pause in between words. In 1971, the Defense Advanced Research Agent Project Agency (DARPA) recognized the importance of speech recognition and established the Speech Understanding Research (SUR) program [16]. This program supported a group at Carnegie Mellon University, led by Raj Reddy, that developed the Harpy Speech Recognition. Other innovations in speech recognition systems created by this group include speaker-independent speech recognition, continuous speech recognition system, and Hearsay, Dragon, Harpy, and Sphinx I/II systems.

Harpy was a machine that had the ability to understand around 1011 words [17]. It was developed after the Hearsay-I system and the Dragon system so that it had the features of the Hearsay-I system and the Dragon system. Hearsay-I was the first successful attempt of continuous speech recognition that was not required to pause in between words. It was the first time speech was modeled as a hidden stochastic process in Dragon systems. Harpy had taken advantage of both systems, thus presenting the new search concept beam of search. A beam search was used for efficient searching and matching [3]. In the following years, many features including speaker-independent speech recognition and a large number of vocabularies were added to Harpy. Sphinx I/II systems could be described as a new version of Harpy [18].

The HMM approach to speech recognition was used by James Baker, who was a student of Raj Reddy at Carnegie Mellon University in 1976. The HMMs are generally used to deal with the variability of speech. While older approaches simply searched sound patterns and phonemes for words, HMM models predicted possible words. HMM became popular in the 1980s, and its popularity continued to increase in the following years. It supports a generic technique that is still used in many multi-languages speech recognition systems. From the 1980s to the 2000s, the following developments in speech recognition occurred;

- Almost all speech recognition systems used HMM as an acoustic model.
- Large-vocabulary, continuous, and speaker-independent systems were designed.
- Microsoft established a speech recognition research group led by Xuedong Huang.
- Commercial speech recognition products were introduced.

From the early 2000s to the present, the following developments were seen and continues to progress:

- Deep learning methods were applied to speech recognition systems, replacing older methods and resulting in tremendous progress in recognition rate. Companies invested in deep learning technologies to provide robust and high-accuracy speech recognition applications. As an example, Microsoft reduced the error rate of their speech recognition by 30% in 2012 [19].

- The major technology companies Google, Microsoft, and IBM provided cloud system application programming interfaces (APIs) that enabled users to instantly translate spoken words to text.
- The use of cloud systems made speech recognition systems available to use in mobile devices, such as Apple's Siri in 2011. Many high-accuracy applications have been developed since then.
- Speech recognition accuracy reached that of human accuracy which is around 5.0%.

In summary, speech recognition has progressed considerably along with recent developments over the past 70 years. In particular, using both deep learning methods and cloud systems have greatly affected these systems, and increased accuracy.

2.1.3 Speech to Text Systems

Generally, speech to text systems can be explained by converting speech signals into meaningful text. Historically, the initial goal in the field of speech recognition was to convert the speech signals to text form with low word error rate. Over the years, the evolution of technology has led to increasing computing power and adopting cloud systems. Thus its application areas have increased. The application areas can be categorized into two major systems: Voice/Speech Command Systems and Automatic Speech Recognition Systems (ASR). Voice/Speech command systems have a wide range of applications. Some of them are Voice Dialling, Robotics, Interactive Voice Response, Aided Communication and Monitoring, and Voice Control Systems.

The ultimate development of speech to text systems are for two basic reasons: The increase in application areas for voice services and significant improvement in speech recognition technologies [20]. As shown above, voice command systems have a wide range of applications and these examples can be increased. A common feature in all applications is converting the speech signals into meaningful text and taking the necessary action by means of text.

The ASR, being the subject of this study, is a speech recognition system that converts speech signals into the corresponding meaningful text without facilitating an appropriate action. ASR systems could be used to see what is spoken on screen instantly. In this thesis, the aim is to convert noisy speech signals to a meaningful text by using noise cancellation and cloud systems.

2.2 Deep Learning for Speech Recognition

Deep learning is one of the research areas of machine learning that is based on learning data representations. It is also known as deep structured learning. It is composed of multiple layers, such that each consecutive layer uses the output from the previous layer as input. Each layer is connected to the previous and the next layer. The layers are called:

1. **Input Layer:** Receives input data and then passes input to the first hidden layer.

2. **Hidden Layer(s):** Compute mathematical operations with the given input. The word *"Deep"* is related to have how many hidden layers are presented.

3. **Output Layer:** Returns the result.

To achieve better results, deep learning systems need very large data set and large computational power. In older algorithms, if the amount of data is increased, the performance also increases to a certain level. Thereafter, it remains constant. In the case of deep learning, the performance continues to increase. Unlike traditional machine learning systems, deep learning systems can handle very large sets of raw data and learn by feeding raw data with representations that are automatically detected or classified by representative learning [21]. These kinds of methods have played an important role in the solution and development of problems that have been going on for many years in speech recognition [22, 23].

The components of basic speech recognition systems are introduced in the previous section. HMM and GMM were used together before deep learning techniques were used in this field. The shortcoming of GMMs is overcome as a result of the advancement in computing power, and the development of machine learning techniques. This has led to the use of deep learning methods, which has become inevitable in speech recognition systems, with the help of Deep Neural Networks (DNNs). The advantages of DNN include:

1. Time for overfitting, fine tuning, and training are reduced.

2. The DNN can handle data representation problem.

3. The use of DNN and HMM improve word recognition rate. This hybrid architecture can efficiently handle very large amount of data by removing uncertainties. Also, this architecture facilitates the use of speech recognition on mobile devices.

2.3 Speech Recognition Using Cloud Computing

The definition of cloud computing basically revolves around: storing, analyzing, and processing of data by connecting remote servers via the internet [24]. It is a new era for computing as it overcomes the limitation of resources [25]. The service provider, such as Amazon Web Services and Microsoft Azure, manages the resources which are based on demand quantities. The number of resources required by the user can change from time to time. The service providers thus need to adjust resources due to the elasticity of cloud-based services. Initially, cloud-based services were used on computers with sufficient internet speed connection. Over the years, advancement in computing power and increasing battery life facilitated the use of cloud computing on mobile devices. Thus, mobile devices became pervasive. Even though there have been many considerable technological advancements in mobile devices, the available applications involve much computation and data. This does not make sense to compute locally on the mobile devices; rather, cloud computing services are used. There is a

novel framework, developed to overcome mobile application constraints, which comes along with a module which recommends a dynamic decision mechanism, whether the application could better be run locally or through the use of cloud services [26]. By adopting the cloud services, it provides offloading, storing and computing data to the cloud, thus saving computation energy and storage.

Speech recognition is one of the most widely used application areas in cloud computing. Nowadays, the great majority of speech recognition applications on mobile devices use the cloud for the recognition task. The major technology companies provide APIs that enable audio signals or its feature vectors to be sent to the cloud server through the internet. Thereafter, their responsibilities would be and waited for. This process basically consists of 3 steps:

1. By sending audio signals from the mobile application to a cloud server.
2. By converting audio signals to meaningful text on the cloud system.
3. By sending the text equivalent results to the mobile application.

The cloud servers not only process and recognize the audio signals, it also determines the intent of the recognized text, by using a large vocabulary dataset. Using cloud computing has an enormous advantage to overcome mobile device constraints. Despite all the advantages, there are some shortcomings that should be considered when using speech recognition systems. They are:

1. **Reliability:** The cloud systems could be used at any time. The computing load may change from time to time. The cloud systems must ensure that they could provide services at any time in any quantity. Most of the cloud systems back up their systems to prevent communication outages.
2. **Privacy and Security:** In cloud computing, all data and computing resources are moving to the cloud. Thus, their privacy and security depend on the cloud system's security measures. The security and privacy problems do often happen and these are the challenges of our time. Big technological companies, even Google and Twitter [27, 28], can not fully solve this problem [25]. In our case, we are assuming that no private data will be used, so these issues are out of our concern.

The cloud computing should have minimum response delay and maximum accuracy in order to make use of it our in daily lives. Due to technological improvements in the past decades, there are many available ASR systems which include Google, Microsoft, and IBM, so on. Since there are many available options, it becomes very difficult to make a choice among them. Since most of the cloud systems operate with low delay, the two essential features we are looking for are noise robustness and low word error rates. By considering these two features, we chose to investigate major three cloud systems: Google, Microsoft, and IBM.

We compared above mentioned ASR systems with a number of different aspects explained in the following sections.

2.3.1 Google

Google has a speech group to develop speech recognition systems which started in 2005 [29]. Since then, the group has been innovating many different speech recognition systems. Some of them are Goog411, Voice Search on Mobile Devices, Voice API for Android Operating System (OS), Youtube Transcription, and Speech Recognition API for Cloud Systems.

Since machine learning and artificial technologies are used for speech recognition systems, these led to significant improvements in WER. Google currently achieved WER of 4.9%, which is the same as the human accuracy and the lowest error rate among the other systems. That is a big improvement since Google achieved 23% in 2013 and 8% in 2015. The secret of this success is the investments made in machine learning and deep learning technologies over the years according to Pichai [30]. Google speech API has the following advantages:

1. It recognizes more than 80 languages and dialects.
2. Multi-audio encodings are supported, including FLAC, AMR, PCMU, and Linear-16 [31].
3. It informs about other possible interpretations of the audio.
4. It uses both remote procedure call (gRPC) and representational state transfer (REST) protocols.

2.3.2 IBM

IBM is one of the well-established technology companies that manufactures mainly computer hardware and software. IBM researchers have been dealing with speech recognition since the 1950s. Since then, IBM has developed many speech recognition products. Some of them are IBM 701, IBM Shoebox, Pioneering Speech Recognition, IBM Via Voice, and IBM Watson.

IBM Watson's WER is 5.5% which is close to human accuracy [32]. It was 43% in 1995, 15.2% in 2004, and 6.9% in 2016. IBM has been advancing developing in deep learning technologies over the years [33]. The technology company has been using different acoustic and language models together to achieve better performance, with an ultimate aim of exposing both acoustic and language models with a very large data set to achieve higher accuracy. IBM Watson has the following advantages:

1. Multi-audio encodings are supported, including WAV, FLAC, and PCM.
2. It recognizes and supports 7 languages [34].
3. It uses both REST and WebSocket protocols.

2.3.3 Microsoft

Microsoft is another technology company that develops mainly software products such as Windows Operating Systems. The company has also involved in speech recognition by hiring top researchers from the Carnegie Mellon University to develop the Sphinx-II speech recognition system in 1993 [35]. This group has continued to grow since then and have developed several speech recognition systems. Some of them are as follows: Microsoft SAPI, Microsoft Voice Command, Microsoft Cortana, and Microsoft Bing.

According to Xuedong Huang, the following three characteristics have enabled the speech technology to reach human accuracy [36].

1. Data: When speech recognition systems are used frequently, more data is collected and the systems get better by learning from those data garnered.
2. Computing Power: Mobile devices are resource-constrained. Cloud computing provides resources for recognition.
3. Machine Learning: When artificial intelligence technologies improved, researchers tried to use DNNs to train systems for better understanding.

Microsoft has made a major progress in speech recognition by adopting DNN and Computational Network Toolkit (CNTK). CNTK provides optimizations in order to run deep learning algorithms much faster [37]. Microsoft Speech Assistant and Cortana uses both CNTK and GPU clusters to ingest more data [37]. Microsoft's current WER is 5.1% which is close to human accuracy [38]. It was 6.3% in 2016 and around 17% four years ago [38]. Microsoft Bing Speech has the following advantages:

1. Multi-audio encodings are supported, including WAV, PCM, and Linear-16.
2. It recognizes and supports around 28 languages.
3. It uses both REST and WebSocket protocols.

2.4 Speech Recognition on Mobile Devices

Mobile devices or Smartphones have been very popular over the last decade. Many vendors produce smartphones that come with advanced computational power and heuristic features. They became popular with the introduction of Apple's iPhone in 2007. In 2017, the number of smartphone users was around 2.32 billion worldwide [39]. Since almost one-third of the world population uses a smartphone, there is undoubtedly stiff competition among vendors to garner customers. The vendors need to provide longer battery life and better computation power, due to resource feasibility of mobile devices. Applications could run locally or in cloud services on mobile devices. Speech recognition is one of the applications that its computation could be offloaded to a remote service such as cloud. Speech recognition could also run locally. Due to the limitation of mobile devices, Apple's Siri prefers running remotely in cloud services. This is achieved by sending its audio or feature vectors to the cloud server

by means of the internet, thereafter a response is waited. The mobile device could be thought of as a client and at the same time, a cloud server. This process basically consists of 3 steps:

- By sending audio from client to cloud server.
- By converting audio to meaningful text.
- By sending equivalent text results to the client.

It is noteworthy to state that these phases should have minimum latency in order to satisfy the real-time performance.

In this study, we used a mobile device for speech recognition. To recognize speech we used the cloud system to decrease the computation power, which in turn results in increasing the battery life.

2.5 Challenges for Applications Using Speech Recognition

Speech recognition has been studied for the past five decades, and it has been used in many different areas, such as voice dialing, web surfing, health care, and many others. Appreciable progress has been recorded from the 1950s, to make robust and speaker independent speech recognition systems. Since the DARPA sponsored the SUR program, WER became the main metric for speech processing evaluation [3]. As of today, the best word error rate is 4.9% which is the same as that of human, as claimed by Google [40]. Google achieved 23% in 2013. As indicated in the numbers, there has been a big improvement. To achieve low error rate with a robust speaker-independence system, the researchers had to overcome some challenges. Theses challenges include speaker dependence, accuracy, latency, noise robustness, and reliability of the system. Each of these challenges are discussed in the following sections.

2.5.1 Speaker Dependence

Speech signals have a large range of variability. Each person has unique sound characteristics such that it is impossible to produce the exact same sound with different people. Even the same person cannot reproduce exactly the same sound when it is attempted [41]. There is always an occurrence of little variations. Environmental conditions should also be taken into consideration.

Variability of speech signals and their handling is the main challenge for the ASR systems. It is possible to get a high accuracy rate for single speaker speech in a quiet environment. However, adding some background noise to the environment, changing speaker, changing microphone or moving microphone position according to the speaker may result in lower accuracy. So, speech recognition designers must take these into account.

For variability, speech recognition systems can be divided into two categories:

- **Speaker Independent (SI) Systems:** They are designed to recognize any speaker's speech. It is necessary to train SI systems with a large number of different people so that they could provide almost the same accuracy for all.

- **Speaker Dependent (SD) Systems:** The SD systems focus on sounds that are produced by specific speakers. They show good performance for the specific speaker, but poor performance for different speakers [42]. They learn speaker's voice characteristics through training using the speaker's voice.

Mostly, old systems were SD systems due to technological limitations. SI systems require more memory and computational power which were absent in initial speech recognition systems. Since the speech recognition's application areas are getting wider, most people use these systems for different purposes. This, however, forces today's speech recognition systems to be speaker independent systems. The aim is to provide the best accuracy independently from the person speaking.

2.5.2 Delay

Delay is another crucial parameter for speech recognition systems. Especially, when cloud-based speech recognition systems are involved, there should be minimum delay due to cloud access through the network. When the delay gets higher, speech recognition systems produce more inconsistent results, increasing the WER and making the system unusable. The performance is aimed to be consistent in all circumstances. Delay can vary under different network conditions. When the network is involved, the following directly affect the speech recognition systems' performance: The packet loss, jitter (i.e., the time variation of received packets), used network protocol, and bandwidth.

The packet loss and jitter have a significant effect on delay [43]. The used network protocol determines whether there will be a packet or not. Due to accuracy most of the cloud systems use Transmission Control Protocol (TCP) connection. TCP is guaranteed for packet reception. However, by using User Datagram Protocol (UDP), round trip time becomes minimized, which is desirable for real-time requirement, but causes poor performance in recognition. Typical bandwidth is around 2 Mbps for 3G connection and around 12 Mbps for 4G connection. Most of the mobile devices use at least 3G connection for their internet access, which is enough to transmit audio through the internet.

2.5.3 Noise and Interference

One of the fundamental challenges of speech recognition systems is noise and speech interference. Noise is present almost in all environments. Its characteristics may vary over time as the environment changes. Every day, people are exposed to more or less amount of noise in almost all environments. Various types of noise that humans could be exposed to are interfering speech and other sounds, traffic noise, crowd noise, machine noise, white Gaussian noise, and so on.

According to an experiment that was conducted in the United States, the majority of the population who are exposed to a noisy environment could be predisposed to hearing problems [44]. Any unwanted audible sound is called noise. Yet, the noise level is an important parameter that affects the extent of hearing problems. It is also important for speech recognition systems.

Normal speech is around 55-65 dB. Prolonged exposure to any sound that is above 80 dB (A) is damaging to the ear and requires intervention. Noise also affects speech intelligibility in daily life. This specifically, affects older people, children, and people who are suffering from hearing problems [45, 46]. Noise reduces people's quality of life. In some situations like military communication, missing even a word would not be acceptable. In the real world, speech communications usually involve multiple speakers and more or less background noise. Since most of the speech recognition systems require a microphone to capture sound waves, the microphone should be placed near the speaker. This, however, might be impossible because from time to time, there is always a certain distance between the speaker and the microphone. In this case, original speech signals are distorted by the reverberation of environment and the speech interference [47]. A classical example is *cocktail party effect* in which a number of people are talking at the same time with background music [48]. In this case, some questions could come to mind:

1. How can speech recognition systems recognize what people are saying?

2. Which speaker should the speech recognition system focus on?

In order to handle these situations, speech recognition systems use many microphones that are directed to a specific person, rather than others [49]. However, the captured speech signals by microphones generally contain additive noise. Noise can degrade the speech recognition systems's accuracy. There are some other factors that could affect the speech recognition performance under noise. They are:

1. **Gender:** Human hearing ranges from 20 to 20000 Hz. The frequency ranges of the voice of male and female are different. Generally, female voices have a higher frequency than male voices. This means that male voices are spread over lower frequency bands which make them vulnerable to background noise, which frequently occupy lower frequencies.

2. **Reverberation:** It is generally explained as the elongation of sound waves as a result of its reflections on surfaces. Speech communication occurs in noisy and reverberant environments. Reverberation causes degradation of speech recognition performance due to the distortion of the original speech signal [50]. To achieve better recognition performance in reverberant environments, SNR should be higher [51].

As a result, noise, speech interference and reverberation are the main factors that directly affect the speech recognition performance. In the following section, noise cancellation methods found in the literature would be delved into.

2.5.4 Reliability of the System

Speech recognition systems must be reliable under all circumstances. Reliability can be described as the ability of the system to keep operating over time and producing the same results. It is indispensable for these systems. As the systems evolve, the results are expected to be better. Nowadays, most of the speech recognition systems use the cloud technology. This means that the whole vocabulary dataset and used algorithms are stored in the cloud systems. So, it is easier to categorically state that the reliability of these kind of systems depends on the cloud systems. Aside from the cloud systems, noise robustness and speaker dependence also directly affect the reliability of the systems.

2.6 Noise Cancellation Methodologies

Today, most of the cloud systems provide speech recognition accuracy which is the same as humans. However, it is not clear how these system's accuracies are tested. The technology companies claim that these systems repel noise. However, despite all these improvements, the success of these systems in a noisy environment is still insufficient. Most of the time, performance tests are conducted under low level noise or in noiseless environments, which poses a challenge to achieving a high success rate in a noisy environment. There are different approaches to overcome noise in a speech signal.

Noise cancellation can be described as removing noise contamination from the speech. As pointed out in Section 2.5.3, speech recognition degrades due to additive noise and reverberation. Moreover, noise characteristics can change from time to time and from place to place. Also, there are different types of noise which was explained in Section 2.5.3. Therefore, its estimation and cancellation is a problem. For these reasons, there is no generally accepted versatile methodology that could be applied for noise cancellation. So, the applied methodologies could change due to noise types and characteristics. We will examine mostly widely used noise cancellation methodologies found in the literature.

- **Generic Noise Cancellation Algorithms:** Noise Cancellation Algorithms eliminate noise from speech signal and increase the SNR while preserving the characteristics of original speech signal. They generally run on a specially designed processor, like Digital Signal Processors (DPSs), due to required high computing power. It is generally assumed that the amplitude of the ambient noise is low. The most commonly used algorithm is spectral subtraction.
  - **Spectral Subtraction:** Spectral subtraction is the most widely used single channel noise removing technique [52]. In this method, the noise is estimated in short pause intervals and subtracted from the speech to increase speech intelligibility [53]. Additive background noise is assumed to be stationary for the estimation of noise in short pauses.
- **Filtering Techniques:** Filtering attempts to eliminate unwanted noise from the original signal by extraction of useful information and preserving the original

signal. There are several filtering techniques. However, all filters do not perform equally. Some of them are:

  - **Kalman Filter:** Kalman filter estimates uncertainties of variables and minimizes the mean square error by observing the signal over time [54]. It is also called as linear quadratic estimation. In speech recognition applications, bidirectional Kalman filter eliminates non-stationary noise from the speech signal by utilizing the previous state. It consists of two steps. The first step, which is prediction, estimates the variables along with their uncertainties. The second step, which is correction, obtains the variables improved by using feedback control [55]. It is a recursive algorithm. It can be used in real-time applications by utilizing the past and the present state information. Thus, no additional memory is required.
  - **Adaptive Filter:** Adaptive filtering technique first analyses the characteristics of the noise and then adjusts itself with estimation error. These two steps work together to feedback the system by modifying coefficients of the applied filters [56]. It is time-dependent because of changing speech signal parameters. Most adaptive filters are digital filters. They are used in many applications such as Telecom systems and digital cameras.

- **Active Noise Cancellation (ANC) Techniques:** ANC is a technique that attempts to attenuate low-frequency noise. Specially designed circuits produce a signal the same frequency as noise, however, only phase flipped by 180 degrees. Thus, noise is neutralized with the generated wave. This technique is mostly used in noise cancelling headphones, to increase audio quality by eliminating low-frequency noise. ANC performs well for lower frequencies and its performance rapidly decreases when the ambient noise level increases [57].

- **Beamforming Techniques:** Beamforming techniques aim to eliminate noise contamination by focusing on the arrival of signal direction using microphone arrays. The beam could be focused on the source signal. Arrays of microphones that consist of more than one microphone are used in beamforming so that unwanted noise, interfering sounds, and reverberation can be eliminated by separating the incoming signals from the others [58]. Since the SNR is usually low, more than one microphone is required to achieve good signal quality, because utilizing several microphones provides better spatial diversity. Beamforming with a microphone array improves speech intelligibility due to the fact that unwanted sounds are rejected.

  The most common approach of beamforming is delay-and-sum method [59]. In this method, input to each channel of array microphone is delayed to achieve time-alignment of the incoming speech signal for constructive addition of waves. Time-aligned inputs are then weighted and summed to focus on the target direction [60]. Thus, any additive noise signal that is misaligned is eliminated. Besides the delay and sum beamforming, filter and sum beamforming is also widely used. A linear filter is applied to each channel of the array microphone and the results are summed.

- **Blind Source Separation (BSS) Techniques:** BSS techniques are used to separate individual signals from their mixture [61]. They do not assume any information in regards to the source of the signal and interferences. Moreover,

they do not require any training stage. Most widely used BSS technique, which is known as the Independent Component Analysis (ICA) assumes that the signals are statistically independent [62]. Moreover, the mixtures are assumed to be instantaneous mixtures, i.e., weighted and summed signals, which does not take into account the effect of reverberation, which results in signals convolved with different room transfer functions and then added.

Trying to achieve the ICA in the frequency domain is an option, so that the convolution in time domain becomes multiplication in the frequency domain and instantaneous mixture assumption can be made. However, in this case, the permutation problem occurs [63]. As a general restriction of ICA, the number of microphones in the array should be the same or more than the number of signals in the mixture, which is known as the determined, or over-determined cases, respectively.

BSS techniques consist of two steps. The first is identification step which determines the number of the independent speech signal and assigns them to a set of parameters. The second is separation step which eliminates the mixture using parameters obtained in the identification step.

To separate mixture signals in the under-determined case, i.e., when the number of microphones in the array is fewer than the signals in the mixture, time-frequency binary masking has been proposed. The masking term refers to filtering in the time-frequency domain. Initially, the Gaussian mixture of mixture speech signals are filtered in the frequency domain [64]. Then, the speech signals are filtered in the time domain to eliminate stronger noise energy, as a result of which the desired speech signals energy remains [65]. This process basically increases speech intelligibility. After these two steps, the speech signals are ready for recognition.

2.7 Evaluation of Noise Cancellation Methodologies

Among the several methods examined above, none of them meets our requirements because the assumptions made in these methods. The advantages and limitations of the methods are given in Table 2.1.

Unlike the standard noise cancellation algorithms, the chosen sound decomposition method, which will be described in Chapter 3 does not impose any limitation on the spectro-temporal characteristics of the noise. In fact, the noise may be another speech signal as in the case of two or more people talking simultaneously. In such a situation which source is the target and which ones are the noise depends on the listener. For these reasons, the assumptions made by many noise cancellation algorithms, such as noise is occupying low frequencies, or noise is additive white Gaussian, etc. [66] are not valid. Similarly, the performance of deep learning-based systems aiming at noisy speech recognition could only be improved for some simple types of noise, other than interfering speech [67]. ICA-based signal separation can not run in real-time and does not perform well in reverberant environments. Beamforming with large arrays can achieve good sound source isolation, however, they are not practical for use with mobile devices. Therefore, we have utilized a sound decomposition methodology specific to the requirements of a mobile system. The detailed explanation of this

Table2.1: Comparison of Noise Cancellation Methods

Method	*Advantages*	*Limitations*
BSS Techniques	No training phase required. No assumption is made about the source of the signal and other interferences.	The number of microphones should be equal to or higher than the number of sources. Sources are assumed to be independent and sparse.
ANC Techniques	Increased noise attenuation.	The noise frequency should be low.
Spectral Subtraction	Easy implementation.	The noise should be stationary.
Kalman Filtering	Provides the estimation quality and the variance of the estimation error. Mostly used in digital platforms.	The states should be Gaussian. Used only in linear systems.
Adaptive Filtering	Computed in real-time.	It can be generally assumed that the amplitude of ambient noise is low.
Beamforming Technique	It can separate the targeted source easily from the mixture using a microphone array.	Separating speech from the noise with a high SNR requires forming narrow beams, which requires the use of several microphones. Furthermore, using multiple microphones with spacing between them results in a large array size, which may not be practical in the case of mobile devices.

system can be found in Section 3.3.

2.8 Audio Transmission to a Mobile Device

Audio transmission is another criteria for verification of the real-time requirement of the system. There were two options; cable and wireless data transmission. Both options provide sufficient data transmission rates for the real-time requirement. The wireless data transmission was chosen due to the following reasons: Flexibility, mobility, low cost, ease of use on mobile devices, and no cable restriction. However, there are some disadvantages with respect to a cabled communication, such as lower reliability and lower data rates [68]. Two solutions come to mind when it comes to wireless data transmission in mobile devices: Bluetooth and WiFi (Wireless Fidelity).

Bluetooth is wireless communication is based on the radio system. It is used for transferring information between two or more devices. It is designed for both short range and low bandwidth communications, such as sound data transferring. It can also be

used in different application areas such as printers, voice transmission between mobile devices, headsets and so on. Bluetooth communication is designed for establishing a personal network between devices, by replacing cable connection [69].

WiFi is also wireless communication, yet, is not based on the radio system. It allows devices to communicate across both the internet and the local wireless network. It is designed for long range and high bandwidth communications, like streaming video via the internet. Since the internet gets involved, the WiFi application area is very wide such as video conferencing, surfing on the web and so on.

The detailed features of both Bluetooth and WiFi are shown below in Figure 2.2.

Standard	*Bluetooth*	*Wifi*
Protocols	IEEE 802.15.1	IEEE 802.11 a/b/g
Frequency Range	2.4 GHz	2.4, 3.6, 5 GHz
Nominal Range	10 - 30 m	10 - 100 m
Power Consumption	Low	High
Max Data Rate	24 Mbps	600 Mbps

Figure 2.2: Comparison of Bluetooth and WiFi

2.9 Evaluation of Speech Recognition Performance

We have explained the speech recognition system and the factors that affect their performance in the previous sections. However, how exactly can we evaluate the performance of these systems? Is it enough to just convert speech to meaningful text form? How can we decide which speech recognition system is better? Two metrics are very useful when evaluating the performance of speech recognition systems: Accuracy and noise robustness.

Accuracy is the first and the most important metric when evaluating the performance of speech recognition. This is because all proposed speech recognition systems are introduced by explaining their accuracy rate. However, it may not always be clear how, i.e., under which conditions the accuracy was tested; especially in a noisy environment. Therefore, noise robustness is another evaluative metric. As stated earlier, most of the speech recognition systems provide poor performance in a noisy environment. Both accuracy and noise robustness will be delved into in subsequent sections.

2.9.1 Accuracy

The accuracy can be described as the closeness of the correctly identified words to the actually spoken words. The accuracy is a key metric for speech recognition systems. Since the early years, the ultimate aim of the researchers has been to obtain the best accuracy for speech recognition. When major technology companies introduce new speech recognition systems, they often brag about having the lowest error rate.

Craig Federighi, for example, who was Apple's senior vice present of software engineering, stated that Apple's Siri is more accurate than Google's in 2015. A number of researchers, engineers, and scientists have been working to improve the accuracy of speech to text conversion. In order to evaluate the performance of speech recognition, the following four metrics can be used: Correct Word Rate (CWR), Word Accuracy Rate (WAR), Sentence Error Rate (SER), and Word Error Rate (WER).

In the following equations, S denotes the total number of substitutions, I denotes the total number of insertions, D denotes the total number of deletions, and N denotes the total number of words to be referenced.

CWR is calculated by the following formula:

$$CWR = \frac{N - D - S}{N} x100 \tag{2.5}$$

WAR is calculated by the following formula:

$$WAR = \frac{N - D - S - I}{N} x100 \tag{2.6}$$

SER is calculated by the following formula:

$$SER = \frac{numberofcorrectsentences}{totalofcorrectsentences} x100 \tag{2.7}$$

WER is calculated by the following formula:

$$WER = \frac{D + S + I}{N} x100 \tag{2.8}$$

Besides the above metrics, there are also some other metrics, such as the Character error and utterance error. However, they are only used for specific purposes. In this study, we use WER metric, which is widely used in the speech recognition domain. Moreover, all speech recognition systems show the WER as a performance parameter. However, WER has some disadvantages such as:

- The result can exceed 100% if the sum of the number of deletions, substitutions and insertions exceed the total number of words.
- It does not tell how successful the system is. It just states that one is better than the other.

Over the years, the researchers, engineers, and scientists have tried to reduce the WER, thus improving the accuracy of speech recognition. WER was around 43% in 1995 as achieved by IBM. It was 15.2% in 2004, and Microsoft achieved 6.3% WER in 2016 [37]. As stated earlier, the contemporary and the best WER is 4.9% as claimed by Google. This is the same error rate as humans. The error rate has been drastically reduced due to a number of breakthroughs in the speech recognition technology, such as the improvement in speech decoding techniques, artificial intelligence, cloud computing and development of devices that have more computational power. This is definitely a big leap in progress. It was not possible to achieve these word error rates sixty years ago. As technology evolves, scientists still continue to improve the accuracy. The ultimate goal is to achieve better speech accuracy than humans.

2.9.2 Noise Robustness

One of the fundamental metrics of speech recognition systems is noise robustness. Noise robustness can be expressed as the speech recognition system's ability to withstand decreases in SNR. Noise is present almost in all environments, thus the captured signals by microphones generally contain speech with additive noise. Noise degrades speech recognition systems' accuracy as well as the speech intelligibility. Noise robustness is essential, since speech recognition systems are generally used in noisy environments, such as a voice assistant on a busy street.

The speech recognition systems' performances vary according to the noise level and its spectro-temporal characteristics. The majority of speech recognition systems work reasonably well in quiet environments, yet produce high error rates in noisy environments. Both academic researchers and engineers from various industries have been working on this problem, i.e., noise robustness [70].

Besides noise, speech interference and reverberation are other factors that leads to the degradation of speech intelligibility. While the speech recognition systems of major technology companies are claimed to process noisy audio data from various media smoothly; most of them are vulnerable to noise, speech interference, and reverberation. When SNR decreases, the performance of speech recognition systems decreases drastically. In this thesis, the performances of major speech recognition systems have been tested with noisy speech to determine their noise robustness, and a system has been proposed to provide reasonable speech recognition accuracy in noisy environments.

CHAPTER 3

METHODOLOGY

In this chapter, we will describe the systems that came to the forefront in the design phase. Their components and subsystems will be explained in detail.

3.1 System Design Overview

In this thesis, we will mainly focus on speech recognition on mobile devices. The design consists of two subsystems: The accessory and the device. The accessory subsystem consists of three sub-parts: Data acquisition, digital signal processing, and Bluetooth transmitter.

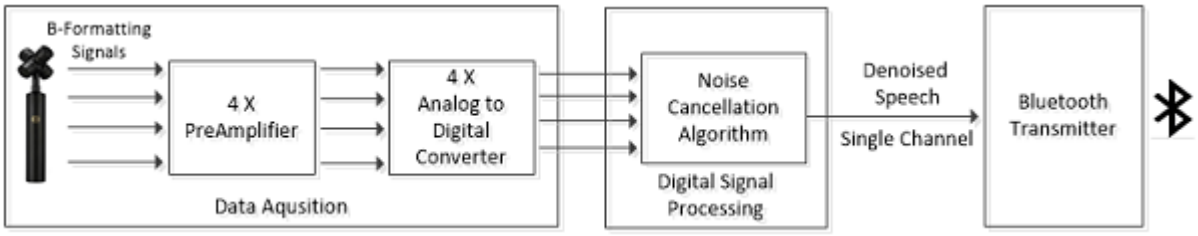


Figure 3.1: The Accessory Subsystem of the Design

Data Acquisition: Data acquisition captures sound signals in the environment via a microphone array. It then amplifies the captured sound signals and converts these analog signals to digital.
Digital Signal Processing (DSP): The DSP executes a sound decomposition algorithm on the digital signals.
Bluetooth Transmitter: Bluetooth transmitter transmits the processed signals to a mobile device through a single channel.

The device subsystem consists of two sub-parts: Mobile application and cloud system.

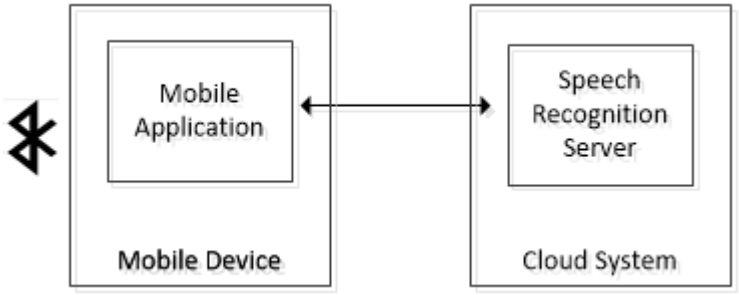


Figure 3.2: The Device Subsystem of the Design

Mobile Application: This application runs on a mobile device that receives the denoised speech signals from the accessory subsystem via Bluetooth. It sends them to the cloud servers and shows the text equivalent of the conversations on the mobile device screen in real time. For the facility of testing within the mobile application, we chose the Android operating system.
Cloud System: The cloud system receives the audio signals via the internet, converts them into a meaningful text and sends them to the mobile device application. For the cloud system, we chose to use Google, since it has the lowest word error rate.

Apart from this design, there could be the option of running the noise cancellation algorithm on the mobile device. In this case, the system would not perform in real time, due to the resource limitation problem in mobile devices. Moreover, transmitting multichannel audio would be necessary, which would complicate the system. Executing the noise cancellation algorithm on the mobile device would also cause the battery to be depleted in a short time. If the application is required to work for a long period, an extra battery would be needed.

In this study, we implemented the whole device subsystem. The accessory subsystem, apart from the Bluetooth Transmitter had already been implemented as part of [71] and was made available for this study. The details of the proposed system are provided in the following sections.

3.2 Data Acquisition

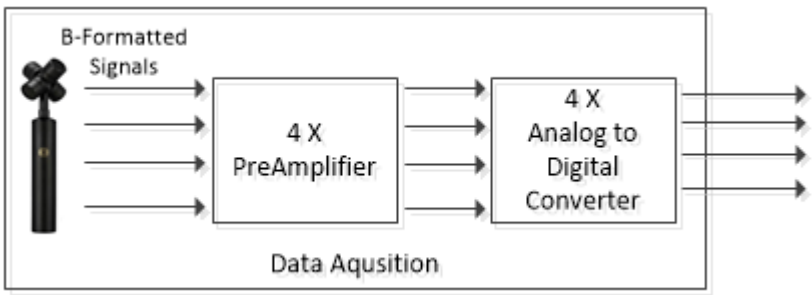


Figure 3.3: Data Acquisition Part of the Design

This part consists of two sub-parts. The first part consists of a tetrahedral microphone array. The second part consists of a preamplifier and an analog to digital converter.

In the designed system, there were several options for the microphone array. We chose to use the tetrahedral microphone array due to its advantages. Determining the location of the speech sources by microphone arrays to carry out beamforming is a problem for speech processing applications [72]. Capturing speech signals from all directions surrounding the microphone is essential for a mobile application. Since Uniform Linear Arrays (ULA)s or other planar geometries could not cover the space in 3D, a 3D symmetric geometry was needed. For a mobile application, a compact size and limited number of microphones is also essential. The key feature of the tetrahedral microphone array is that the microphones are closer to each other. In this way, a compact structure is provided to be compatible with mobile devices. It allows the reception of four-channels of audio signals in such a format that the pressure $p_W(t)$ as well as the pressure gradients on the x, y, z-axes, $p_X(t), p_Y(t), p_Z(t)$, respectively, can be obtained [73].

The preamplifier receives four channel audio signals, increases the voltage to a level sufficient to make it compatible with the analog to digital converter. The analog to digital converter receives four channel pre-amplified audio signals and converts them to the digital form with 44.1 kHz sampling frequency and 16 bits. Thereafter, it transmits them to the digital signal processing unit.

Note that there may be more than one speaker at the same time in the environment and which one is noise depends on the context and the listener. Therefore, it is not possible to perform an automatic analysis on the captured sound signals. For this reason, it is expected that the listener should turn the microphone array's front to the desired position. In this way, the target sound will always be at 0°with respect to the microphone array. Since the microphone array has a three-dimensional symmetrical structure, this rotation could also be done digitally. However, this would still require input from the listener via a separate component, such as a rotary switch, a touchpad like device, etc. which would be connected to the DSP.

In order to test this design we needed large data sets to determine its accuracy and robustness under noise. Therefore, we used prerecorded audio files and simulated the signals that would be obtained with this design. The details of the prerecorded audio files are explained under Section 4.1.

3.3 Noise Cancellation Algorithm Specifications

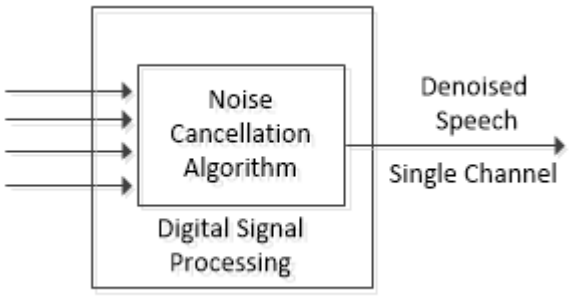


Figure 3.4: Noise Cancellation Algorithm Part of the Design

In this part, the received four-channel digital audio signals are processed on the DSP device in real-time. The DSP receives four-channel audio signals applies noise cancellation algorithm on it and sends the clean audio signal to the Bluetooth transmitter.

This algorithm based on blind source separation technique (see Section 2.6). The real-time sound decomposition with the obtained signals is performed by an acoustic vector processing technique, based on spatial filtering method by determining the direction of the acoustic intensity vector [71]. In this method, the signals showing the pressure and pressure gradients are first converted to the time-frequency domain. For each time-frequency range, the direction of the active intensity vector lies in parallel to the direction of the sound source.

$$\gamma\left(\omega,t\right) = tan^{-1}\frac{Re\{p_W^*(\omega,t)\ p_Y(\omega,t)\}}{Re\{p_W^*(\omega,t)\ p_X(\omega,t)\}} \quad (3.1)$$

Re{ }, denotes the computation of the real value, * denotes the complex conjugate, ω denotes angular frequency, and t denotes for time. The noise-free speech signal is obtained by filtering $s(\omega,t)$, the pressure signal with a spatial filter $f(\gamma(\omega,t);\mu,\kappa)$ oriented in the speech direction:

$$\mathrm{s}(\omega,t) = p_W(\omega,t) f(\gamma(\omega,t);\mu,\kappa) \quad (3.2)$$

This function is chosen for the spatial filter used since the directions of acoustic intensity vectors calculated for a sound source in the presence of reclections resembles the von Mises distribution [74], which is the circular equivalent of the Gauss function:

$$\mathrm{f}(\theta;\mu,\kappa) = \frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)} \quad (3.3)$$

Here, the κ concentration parameter, the μ source direction angle, and the θ current direction angle; $I_0(\kappa)$ denotes the modified Bessel function in the zeroth order. In this method, it is assumed that the directions of the target sounds are known, or the directions of the sounds are first determined [73]. Thereafter, the filtering can be done for these directions. For speech recognition, it is sufficient to specify only the target direction.

3.4 Transfer Media Selection

The detailed explanation of transfer media options was provided in Section 2.8. Both Bluetooth and WiFi provide wireless communication by using radio signals. However, the basic distinction is their purpose of design. While Bluetooth is essentially designed for short range and low power consumption; conversely, WiFi provides long range and high power consumption. By considering data rate, both are sufficient for transferring audio signals. These two options do not have much superiority over each other, and both can be used in our system. Bluetooth has been chosen because the application needs to run for a long time and the power consumption of Bluetooth outperforms that of WiFi.

3.5 Mobile Platform Speech Recognition Application

This part of the system consists of two parts: Mobile Application and Cloud System. The mobile application runs locally on a mobile device, while the cloud system runs on the cloud server. These two systems share information through an API provided by the cloud system.

These systems were examined in detail regarding the Speech Recognition Using Cloud Computing in Section 2.3. The WER is the main performance metric for speech recognition systems and there was a competition between systems to achieve the lowest word error rate on the first trial [3]. All are easy to use and can be deployed on mobile devices; however, it was observed that the Google API has the lowest word error rate among them. Initially, we performed many tests among the three of them (see Section 4.3.1). It can be inferred from the results that Google provides better accuracy. So, Google was be used in our system.

After choosing the cloud system, we needed to select an operating system for mobile application development. Google provides an API for both Android and iOS operating systems, which together are used by 99.6% of new smartphones [75]. 74.2% of all mobile phones worldwide use the Android operating system as of January of 2018 [76], indicating that Android is significantly more widely used than iOS. There are other reasons why Android is preferable to iOS, as follows:

1. The mobile application is easier to test within Android than within iOS.
2. Android is open source.
3. Android is more flexible than iOS.

Due to the reasons stated above, we chose to use the Android operating system for developing an application in order to reach a wider audience.

CHAPTER 4

PERFORMANCE ANALYSIS

In this chapter, we will explain the experimental setup and cover the factors affecting speech recognition. Also, detailed information about the performed tests and the obtained results will be presented.

4.1 The Experimental Setup

The ASR is a technique that is used to automatically convert speech signals into corresponding meaningful text. In order to design the proposed system, we need to ensure that it provides sufficient results in more difficult acoustic environments. Toward that end, a variety of tests were performed using very large datasets. Different audio files were selected from various sources. We conducted the following tests with the prepared setup: Context independent tests, context independent rhyme tests, context independent tests with different SNRs, and context dependent tests.

Although the design proposed in Section 3 was implemented partially (i.e., apart from the Blueetooth tranmission, which was replaced with an audio cable) and was working in real-time, it was not suitable for carrying out detailed tests. In order to create the various audio files used for tests, the audio signals that would be obtained from the proposed design were simulated using real impulse response recordings made with the tetrahedral microphone array of the design, convolving them with clean speech signals and applying the noise cancellation algorithm. In this way, the speech signals and test conditions could be varied and a large set of audio files could be generated.

The characteristics of the audio files of the data sets we use in the tests are given below. For context independent test, we used prerecorded mono audio files that contain a full sentence. The sampling rate was 44100 Hz. The audio files consisted of 1200 different speech samples that were obtained from 300 words spoken in English by four different speakers, two females and two males. The samples were developed by the Public Safety Communications Research Group (PSCR) [77]. The following is an example of the speech sentences; *Please select the word went.* Each sentence begins with the phrase *Please select the word.* Only the last word of the sentence changes. A total of 4800 samples were generated by the following methods:

1. Mixing 1200 speech samples with restaurant noise in which the target sources were at 30°and the restaurant noise was at 130°.

2. Applying noise cancellation algorithms with three different window sizes to the mixed samples. Window sizes were 512, 1024, and 2048, respectively. A shorter window size increases the separation performance, but also increases the artefacts.

Tests were conducted on a total of 6000 sound samples, consisting of 4800 generated sound samples and 1200 original sounds. The entire data set was tested using Google and IBM Watson cloud systems. For the Microsoft Bing cloud system, only clean audio files were tested.

For the context independent rhyme test, we used the same prerecorded mono audio files that were used in the context independent test. The only difference is that, instead of the whole sentences, only the last words of the sentences (the words that were changed) were tested. The reason for this is that the modified rhyme test measures speech intelligibility independent of the content [78,79]. We only used the Google cloud system because it provides better accuracy than the other cloud systems. Unlike the modified rhyme test procedure used to evaluate the performance of ASR technique, all rhymed words were tested separately, and no options were presented. In analyzing the performance of the ASR technique, it is essential to conduct the evaluation independent of the context. Otherwise, it may be possible to estimate a word that would normally not be understood from the context.

For the context independent test with different SNRs, we used the partial prerecorded mono audio files that were used for the context independent test. We produced different speech samples from the dataset. The audio files consist of 100 different speech samples obtained from 50 words spoken by two different speakers, one female, and one male. A total of 1100 samples were generated by adding noise to the original speech samples and then cleaning them by applying the noise cancellation algorithm. The input SNRs for the samples were 0, 3, 6, and 9 dB. The tests were conducted on a total of 1200 sound samples containing 1100 generated sound samples and 100 original sounds. This test measured how robust the designed system was to changes in the noise level. Again, only the Google cloud system was used for this test.

For the context dependent test, we used both prerecorded mono files and stereo audio files that contain full sentences. The original speech samples consisted of 10 different sentences in English obtained from 20 different speakers, 10 females, and 10 males. The contents of these audio files are given in the Appendix B. The sampling rate was 44100 Hz and the files had 16-bit resolution. Initially, we combined all the sentences of one speaker into a single audio file and obtained 200 original clean speech samples. Then the noisy recordings were obtained by mixing female speech and male speech. The sample from the first female speaker was mixed with the sample from the first male speaker; the sample from the second female speaker was mixed with the sample from the second male speaker, etc. While mixing, the female speaker was positioned at 0°and the male speaker's position was changed to 30°, 60°, 120°, and 180°with respect to the microphone array. This positioning was achieved by convolving the speech samples with the room impulse responses recorded when the sound source was at these positions in a listening room with a reverberation time of 0.32 s. A total of 1600 samples were generated with a mixture of four different angles, 30°, 60°, 90°, and 180°, respectively Tests were made on a total of 1800 sound samples that consist of 1600 generated sound samples and 200 original sounds. Again, only the Google cloud

system was used for this test.

Since the datasets used in this study were very large and diversified, it took a long time to record, process, and analyze them. Therefore, we prepared a testing routine that included three applications that run on a computer to reduce the amount of testing time:

1. Audio Reader
2. Output Collector
3. Speech Recognition Performance Evaluator

Audio Reader: This application basically reads audio files from the database and sends the information to cloud servers using an API that was provided by the cloud systems. The cloud systems send back meaningful text that is equivalent to the audio content. Finally, corresponding texts are stored in the database. A different application was developed for each cloud system. The applications were developed by using Java language for Google and IBM and C# language for Microsoft.

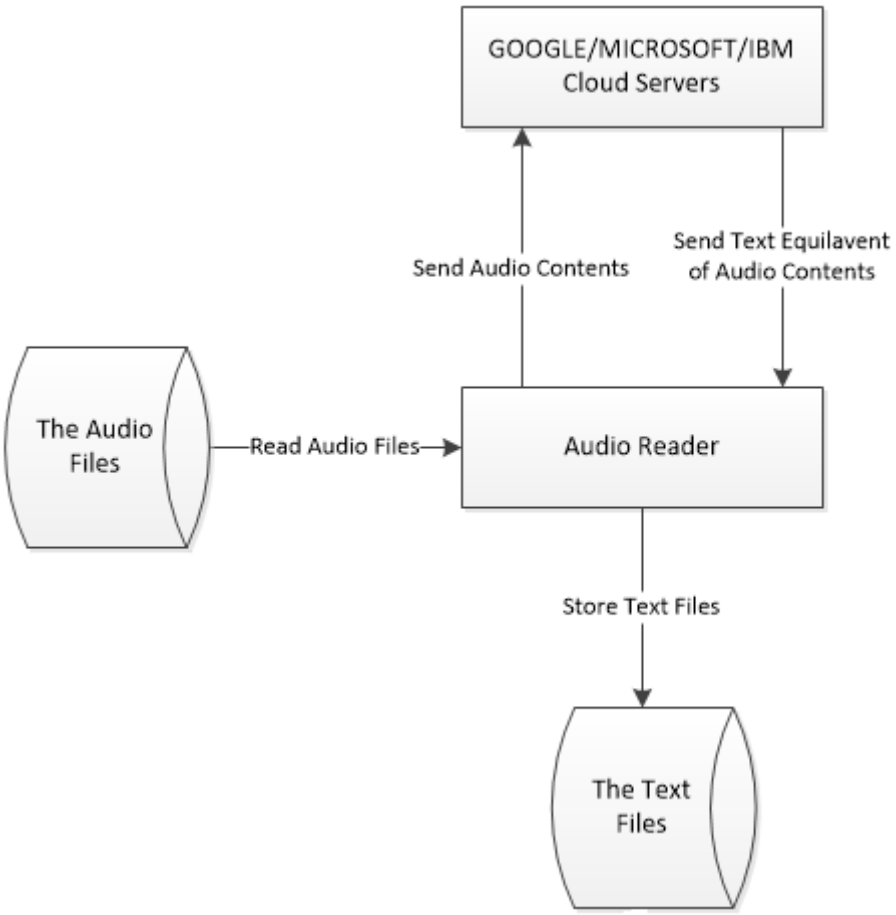


Figure 4.1: Reading the Audio Files and Converting them to Text

Output Collector: It reads all the text files that were stored by the audio reader application and it collects them into a single file based on the audio source. These

data are required for further analysis because we need to determine the robustness and word error rate changes due to the audio source. This application was developed using Java language.

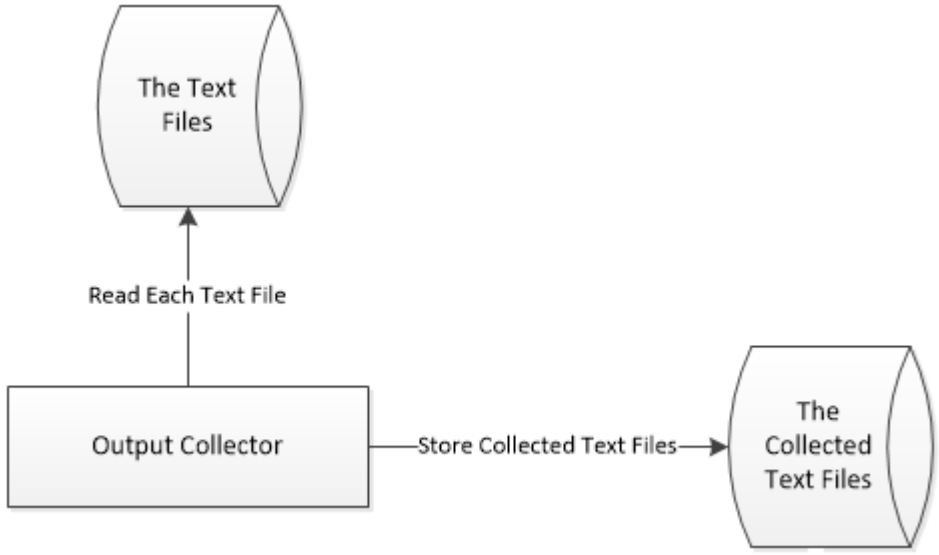


Figure 4.2: Reading the Audio Files and Collecting the Text Files into a Single File

Evaluation of Speech Recognition Performance: It reads the output of the audio equivalent text files and corrects the text files; it then calculates the WER using the Levenshtein Distance algorithm [80]. There are multiple implementation fields of Levenshtein Distance algorithm. Speech recognition implementation is used in this study. Speech recognition implementation of Levenshtein Distance algorithm is a metric to calculate the difference rate of two given strings. The output 0.00% indicates that two given strings are identical. Note that the similarity rate can be higher than 100.0%. This application was developed using MATLAB environment.

4.2 The Covered Speech Recognition Factors

The speech recognition challenges were explained in detail in Section 2.5. We conducted a variety of tests using very large data sets to measure the performance of the designed system. While designing the test procedures, the following issues were considered:

- In all the tests, an equal number of female and male were evaluated using an equal number of speech samples. In addition, the prerecorded speech samples of at least 12 different female and male speakers were tested. These procedures were executed to measure both the impact of the speakers' voices and the speaker's gender on the recognition performance.
- In context dependent rhyme tests, we used prerecorded audio files that have different SNRs to show the speech recognition performance changes based on

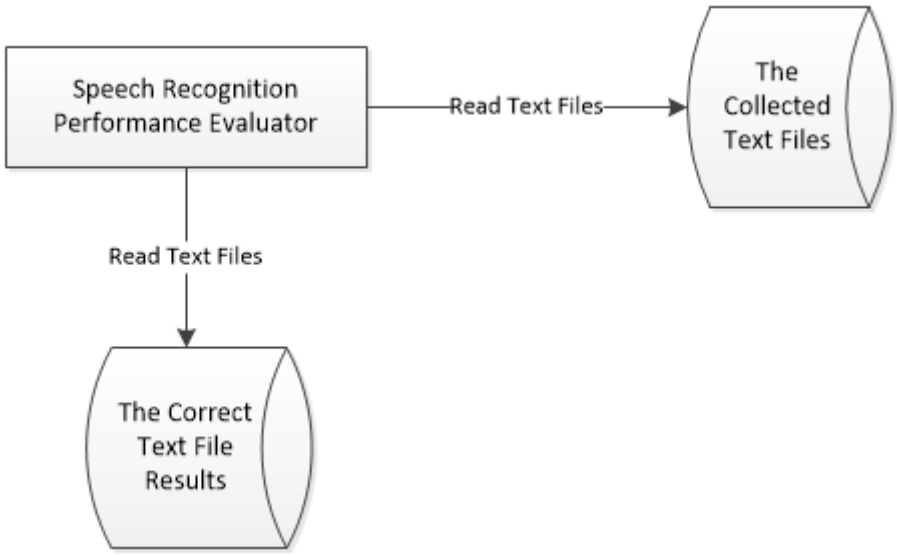


Figure 4.3: The Evaluation of Speech Recognition Performance

the noise level. The following SNRs were used: 0, 3, 6, and 9 dB.

- In the context independent tests, we used three major cloud systems: Google, IBM Watson, and Microsoft to test the prerecorded audio files. Different systems were used to measure the recognition performance of the systems and to also observe which of the three systems outperformed the others. In addition, these systems were used to examine how they behave in relation to noisy audio files.

- In the context independent test, we used separated and mixed speech samples from females and males to test how the designed system handles different sound characteristics of female and male voices. In addition, we also had the opportunity to compare intra-gender speech recognition performances, i.e., among different male speakers as well as among different female speakers.

The designed test procedures were explained in detail above. By considering these procedures, the factors covered can be summarized as noise, speaker dependence and speech interference.

Although the covered factors mentioned above were evaluated, no separate tests were designed to determine the system latency. Among the cloud systems considered, only Google cloud provides a feature to measure the speech-to-text conversion latency time. Using this feature it was found that the latency varies between 20-35 ms. This takes into account the time that elapses through the network, which is known as the response time for cloud systems. In fact, all of the tested cloud systems stated that conversion of spoken audio to text occurred in real time. However, the speech rate should also be considered. In all the test samples, the speaker's speech rate was average and instant conversion may not be possible for quicker speech.

We could not evaluate the cloud system data privacy and security, either, which are also important for our targeted applications. However, it is known that each cloud system has its own rules to protect data. For example, the data is always accessed via

encrypted channels and stored with at least 256-bit Advanced Encryption Standard (AES-256). Thus, no data can be stolen in accordance with these rules.

4.3 Results

In order to evaluate the performance of the designed system, a variety of tests were conducted using very large datasets. The test results are presented separately for each different test. An overview of the results follows immediately. For the reported WERs, 0.00 corresponds to 0.00% and 1.00 corresponds to 100.00%.

4.3.1 Context Independent Test Results

The WERs for context independent tests using Google, IBM Watson, and Microsoft Bing Cloud Systems are given in Tables 4.1, 4.2, and 4.3, respectively. In the tables, *NCA 1*, *NCA 2*, and *NCA 3* represent the noise cancellation algorithms applied with window sizes of 512, 1024, and 2048 samples, respectively. *NO*, means no output, i.e., the system did not provide any output. *Clean* means that there is no noise in the audio files. *Noisy* means that audio files were mixed with restaurant noise. All audio files were tested with Google and IBM Watson cloud systems. Only the clean audio files were tested with Microsoft Bing cloud system, as it gave higher WERs even for clean audio files. So, it was unnecessary to test its performance under noise.

Table4.1: WERs for Context Independent Tests Using Google Cloud System

Speaker	*Clean*	*Noisy*	*NCA 1*	*NCA 2*	*NCA 3*
F1	0.0401	1.9956	0.1417	0.1774	0.1694
F2	0.0497	1.8671	0.1047	0.1165	0.1212
M1	0.0778	4.4653	0.2828	0.2815	0.3597
M2	0.0497	4.1138	0.1115	0.1456	0.1231
Average	0.0543	3.1104	0.1601	0.1802	0.1933

Table4.2: WERs for Context Independent Tests Using IBM Watson Cloud System

Speaker	*Clean*	*Noisy*	*NCA 1*	*NCA 2*	*NCA 3*
F1	0.3874	NO	0.8127	0.7747	0.8256
F2	0.3648	NO	0.8072	0.7955	0.7804
M1	0.4245	NO	1.2367	1.3035	1.0748
M2	0.3641	NO	1.0479	1.0310	0.9702
Average	0.3852	NO	0.9761	0.9762	0.9127

For the results obtained with Google cloud system:

- Female WERs are distinctly lower than male WERs in a noisy environment.
- In clean and NCA-applied audio files, the average female WERs are lower than the average male WERs.

Table4.3: WERs for Context Independent Tests Using Microsoft Bing Cloud System

Speaker	*Clean*	*Noisy*	*NCA 1*	*NCA 2*	*NCA 3*
F1	0.3658	NO	NO	NO	NO
F2	0.2759	NO	NO	NO	NO
M1	0.8217	NO	NO	NO	NO
M2	0.3106	NO	NO	NO	NO
Average	0.4435	NO	NO	NO	NO

- The lowest WERs are recorded for clean audio files.
- The speaker M1 has the highest WER of 446.53% achieved among the noisy audio files, and F1 has the lowest WER of 4.01% achieved among the clean audio files.

For the results obtained with IBM cloud system:

- The cloud system did not provide any output for noisy audio files, which was unexpected.
- Female WERs are slightly lower than male WERs in clean audio files.
- The average WERs for females are much lower than the average WERs for males.
- The lowest WERs are achieved for clean audio files.
- The speaker M1 has the highest WER of 123.67% when NCA 1 was applied to the audio files, and M2 has the lowest WER of 36.41% achieved among the clean audio files.

For the results obtained with Microsoft cloud system:

- Only clean audio files were tested.
- The WERs for females are lower than the average WERs for males.
- The speaker M1 has the highest WER of 82.17% achieved, and F2 has the lowest WER of 27.59% achieved.

Among the clean audio files, for Google cloud, the average WER performance was 5.43% which is 7 times better than the performance of IBM Watson (38.52%), and 8 times better than the performance of Microsoft Bing (44.35%). By considering NCA-applied results, the average WERs in Google cloud were 16.01%, 18.02%, and 19.33% when window sizes are 512, 1024, and 2048, respectively. However, average WERs for IBM Watson were 97.61%, 97.62%, and 91.33%, respectively. As observed, Google cloud clearly provides a lower error rate than IBM, which we expected. When window sizes become larger, Google WER increases, but IBM WER decreases.

In the results obtained from Microsoft and IBM Watson cloud systems, different WERs were obtained when noise cancellation algorithms with different window sizes were

applied. In addition, all of the obtained WERs are better than the WERs from the noisy audio files. After applying the algorithms, NCA 1, NCA 2, and NCA 3, the results were 19.42, 17.26, and 16.08 times better, respectively, than the results obtained for noisy audio files in the Google cloud system on average. A shorter window size improves the separation performance, but also increases the level of artifacts. Since the test results show that the shorter window size performs better, it can be said that the level of artifacts does not adversely affect the recognition performance. In the IBM cloud system, the improvement cannot be calculated, because no output could be obtained for the noisy audio files. However, it can only stated NCA3 provides better results among them.

As seen from the information presented in Tables 4.1 and 4.2, Google cloud has the lowest WERs among the three tested cloud systems. We had investigated Google cloud, IBM Watson, and Microsoft Bing cloud systems in Section 2.3. The Google cloud system was reported to have the lowest WER among the other tested cloud systems, which we verified. In addition, we showed that the Google cloud can produce an output even for noisy speech. Since our study aims to measure how robust the designed system is in noisy environments and quantify the improvement achieved with noise cancellation, we need to get output for noisy speech, too. Since only Google cloud provided an output for noisy speech, only it was used in all of the subsequent tests. The used Google cloud speech recognition jar version is 0.34.0-alpha and accessed February 2018.

4.3.2 Context Independent Rhyme Test Results

The WERs for context independent rhyme tests are given in Table 4.4. In addition, the results for different word groups are given in Tables 4.5 and 4.6. There were 50 word groups, and each word group consisted of six rhyming words. The word lists in these word groups are given in Appendix A. In the tables, the following abbreviations are used:

- WGN: Word Group Number
- NCA1F1: Noise Cancellation Algorithm 1 with Female Speaker Number 1
- NCA1M1: Noise Cancellation Algorithm 1 with Male Speaker Number 1
- NCA1F2: Noise Cancellation Algorithm 1 with Female Speaker Number 2
- NCA1M2: Noise Cancellation Algorithm 1 with Male Speaker Number 2
- CF1: Clean Female Speaker Number 1
- CF2: Clean Female Speaker Number 2
- CM1: Clean Male Speaker Number 1
- CM2: Clean Male Speaker Number 2

NCA 1, *NCA 2*, and *NCA 3* represent the noise cancellation algorithms applied with window sizes of 512, 1024, and 2048 samples, respectively. *Clean* means that there is

no noise in the audio files. *Noisy* means that audio files were mixed with restaurant noise.

Table4.4: WERs for Independent Rhyme Tests

Speaker	*Clean*	*Noisy*	*NCA 1*	*NCA 2*	*NCA 3*
F1	0.1867	0.8533	0.3367	0.3767	0.3967
F2	0.2433	0.9400	0.4433	0.4567	0.4600
M1	0.2650	0.9957	0.7267	0.7167	0.7300
M2	0.2300	0.9867	0.4633	0.5300	0.4967
Average	0.2312	0.9440	0.4925	0.5200	0.5208

Table4.5: WERs for Independent Tests for First 25 Word Groups

WGN	*NCA1F1*	*NCA1F2*	*NCA1M1*	*NCA1M2*	*CF1*	*CF2*	*CM1*	*CM2*
1	0.33	0.33	0.83	0.50	0.16	0.16	0.33	0.16
2	0.33	0.00	1.00	0.66	0.16	0.00	0.00	0.00
3	0.00	0.33	0.16	0.33	0.00	0.16	0.33	0.00
4	0.33	0.33	0.50	0.16	0.00	0.00	0.33	0.16
5	0.00	0.33	0.66	0.50	0.16	0.16	0.00	0.33
6	0.50	0.83	0.66	0.00	0.16	0.50	0.00	0.00
7	0.33	0.50	0.50	0.50	0.00	0.00	0.00	0.00
8	0.16	0.33	0.83	0.50	0.33	0.16	0.50	0.33
9	0.50	0.50	1.00	0.16	0.33	0.50	0.33	0.33
10	0.83	0.50	0.83	0.66	0.16	0.50	0.50	0.66
11	0.33	1.00	0.83	0.83	0.33	0.50	0.50	0.66
12	0.66	0.66	1.00	0.83	0.50	0.83	0.50	0.33
13	0.66	0.50	1.00	0.66	0.16	0.50	0.33	0.50
14	0.16	0.33	1.00	0.33	0.16	0.33	0.50	0.33
15	0.00	0.50	0.66	0.16	0.00	0.33	0.00	0.16
16	0.33	0.16	1.00	0.16	0.33	0.16	0.16	0.16
17	0.50	0.50	0.50	0.33	0.50	0.33	0.33	0.33
18	0.33	0.50	0.50	0.00	0.00	0.16	0.33	0.16
19	0.16	0.33	0.83	0.33	0.00	0.00	0.33	0.16
20	0.33	0.33	0.66	0.33	0.16	0.33	0.16	0.33
21	0.50	0.50	0.50	0.66	0.00	0.00	0.00	0.16
22	0.16	0.16	0.50	0.33	0.00	0.00	0.00	0.00
23	0.16	0.00	1.00	0.33	0.00	0.00	0.00	0.00
24	0.16	0.33	0.83	0.50	0.00	0.00	0.33	0.00
25	0.33	0.33	0.83	0.83	0.16	0.00	0.16	0.16

As seen in Table 4.4, because only one word was tested, the error rate varies from 0.00% to 100.00%, unlike in the other tests. The lowest WERs were achieved for clean audio files. The average WERs for females is lower than the average WERs for males. However, when we examine them separately, in some cases the WERs for males is lower than the WERs for females. The applied algorithms NCA 1, NCA 2, and NCA 3 provide an average of 1.91, 1.81, and 1.81 times improvement, respectively, when compared with noisy audio files for the Google cloud. Shorter window sizes increase the separation performance, but also increase the artifacts. However, test results show

Table4.6: WERs for Context Independent Tests for Next 25 Word Groups

WGN	*NCA1F1*	*NCA1F2*	*NCA1M1*	*NCA1M2*	*CF1*	*CF2*	*CM1*	*CM2*
26	0.50	0.16	0.66	0.50	0.16	0.16	0.00	0.33
27	0.50	0.33	1.00	0.66	0.66	0.33	0.50	0.33
28	0.00	0.66	0.83	0.33	0.00	0.00	0.00	0.00
29	0.66	0.50	0.83	0.66	0.33	0.16	0.33	0.16
30	0.33	0.00	0.66	0.50	0.00	0.00	0.16	0.00
31	0.00	0.33	1.00	0.83	0.16	0.33	0.83	0.50
32	0.16	0.66	0.83	0.83	0.16	0.16	0.33	0.50
33	0.50	0.33	1.00	1.00	0.16	0.16	0.16	0.16
34	0.33	0.33	0.66	0.50	0.16	0.33	0.00	0.16
35	0.50	0.50	0.83	0.50	0.33	0.33	0.66	0.33
36	0.16	0.50	0.83	0.33	0.00	0.16	0.16	0.00
37	0.16	0.33	0.50	0.33	0.16	0.16	0.16	0.33
38	0.50	0.83	1.00	0.50	0.33	0.33	0.50	0.50
39	0.33	0.83	0.83	0.66	0.16	0.16	0.50	0.16
40	0.00	0.33	0.50	0.00	0.16	0.16	0.33	0.00
41	0.50	0.50	0.83	0.16	0.33	0.16	0.16	0.16
42	0.33	0.66	0.33	0.50	0.00	0.33	0.00	0.16
43	0.16	0.16	0.50	0.50	0.00	0.16	0.00	0.16
44	0.16	0.83	0.83	0.33	0.33	0.33	0.33	0.16
45	0.50	0.50	0.50	0.50	0.33	0.50	0.33	0.50
46	0.50	0.16	0.66	0.66	0.16	0.16	0.33	0.33
47	0.83	0.83	0.66	0.83	0.66	0.66	0.83	0.66
48	0.16	0.66	0.50	0.33	0.16	0.33	0.00	0.00
49	0.50	0.66	0.16	0.16	0.33	0.50	0.50	0.16
50	0.33	0.33	0.66	0.33	0.16	0.33	0.16	0.16

that the shorter window size performs better for Google cloud.

We think cloud systems may recognize certain words better because of their length or in general their content. To better analyze this situation, we examined 50 rhyming word groups separately. As seen Tables 4.5 and 4.6, for clean audio files, the average score for WGN 7, 22, and 23 was 0.00%, whereas the average score for WGN 12 was 54.00%, which is the highest WER among clean audio files. In NCA1, WGN 7 had the lowest average WER score at 29.00%, whereas WGN 12 had the highest average WER score at 78.75%. WGN 7 consists of the following words; *teak*, *team*, *teal*, *teach*, *tear* and *tease*. WGN 12 consists of the following words; *sum*, *sun*, *sung*, *sup*, *sub*, and *sud*. Our objective in doing this test was to be able to determine if speech-recognition systems recognize certain word groups better than others. Results show it is indeed the case.

4.3.3 Context Independent Tests with Different SNR

The WERs for the context independent tests with different SNRs are presented in Table 4.7. Speech samples from one female speaker and one male speaker were used in the context independent tests, denoted by F1 and M1, respectively in Table 4.7.

The following SNRs were used: 0, 3, 6, and 9dB.

Table4.7: Context Independent Tests with Different SNRs

Speaker	*Audio Type*	*0dB SNR*	*3dB SNR*	*6dB SNR*	*9dB SNR*
F1	Clean	0.028	0.028	0.028	0.028
	Noisy	1.1312	0.295	0.105	0.04
	Noise Cancelled	0.139	0.050	0.034	0.034
M1	Clean	0.044	0.044	0.044	0.044
	Noisy	4.0043	0.7627	0.1597	0.083
	Noise Cancelled	0.1753	0.1493	0.092	0.068

The results clearly indicate that as SNR increases, WER decreases. As SNR increases from 0dB to 3dB, 6dB, and 9 dB, respectively, the noise cancellation algorithm was applied to the audio files, the improvement rates increased 8.13, 5.90, 3.08 and 1.17 times for the female speaker and 22.84, 5.10, 1.73 and 1.22 times for the male speaker. Figure 4.4 and Figure 4.5 show the WERs against different SNRs for the female and male speakers, respectively.

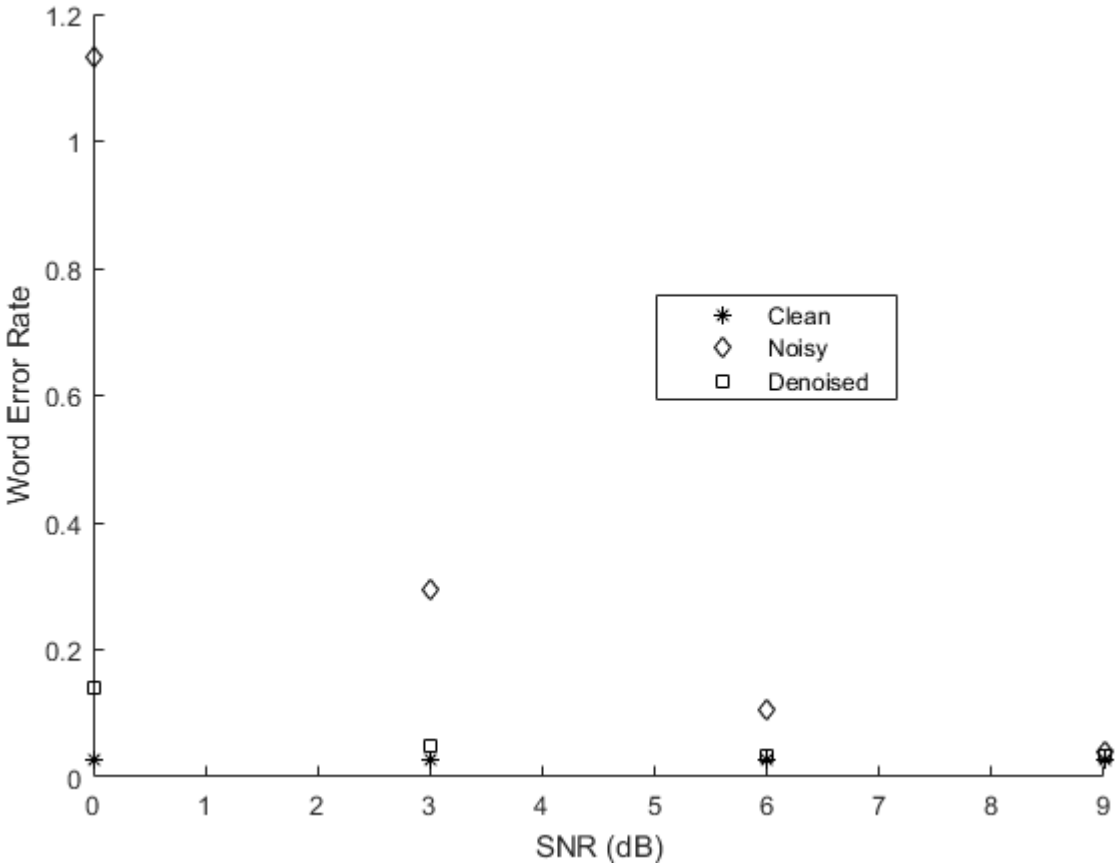


Figure 4.4: Context Independent Tests with Different SNR Results for the Female Speaker

As the SNR increases from 0 dB, WER decreases rapidly for a certain period and

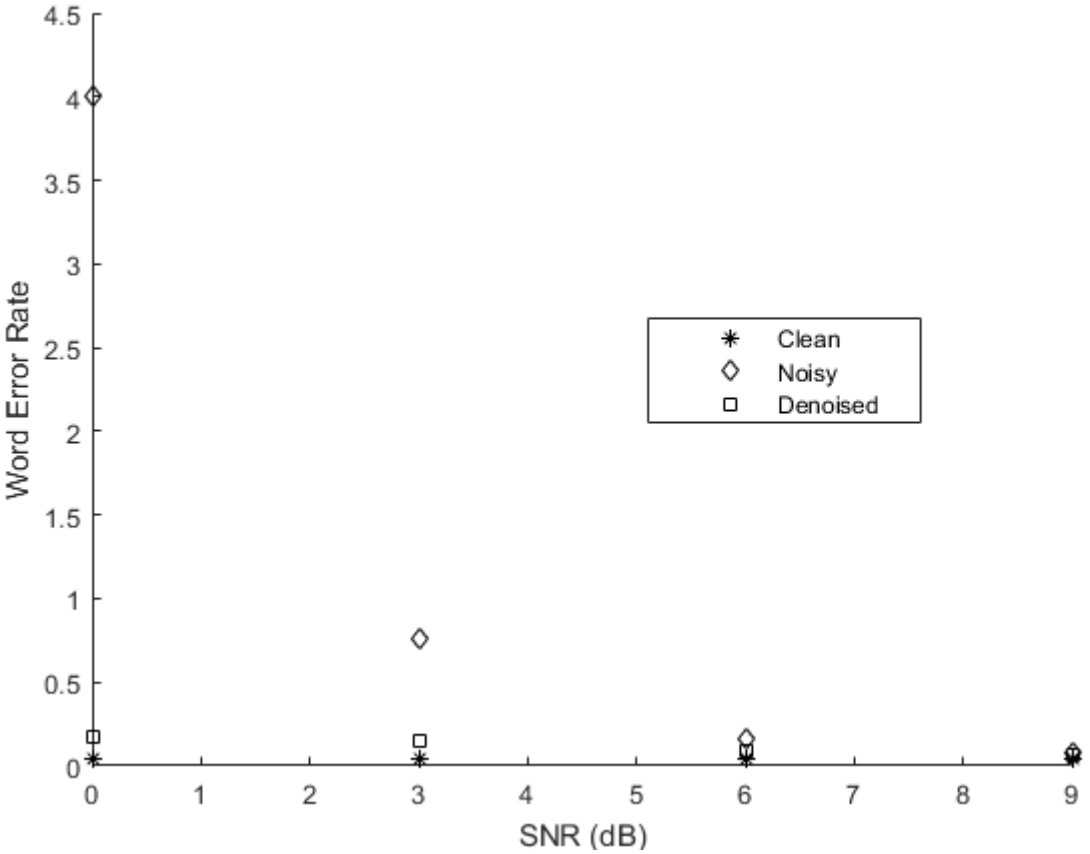


Figure 4.5: Context Independent Tests with Different SNR Results for the Male Speaker

then the rate of change decreases. We predict that the change in WER will be limited when the SNR is equal to or greater than 9dB. When the information presented in Figure 4.4 and Figure 4.5 is examined together,it can be seen that the recognition performance for female speaker is better than the performance for male speaker in all of the tests. This could be due to the fact that frequency bands occupied by male voices are more masked by noise. The spectrums of the original, noisy, and noise cancelled signals whose SNRs are 3 dB are given in Appendix C.

4.3.4 Context Dependent Test Results

The following tests were carried out within the scope of content dependent tests:

1. Individual tests for female speakers
2. Individual Tests for male speakers
3. Context dependent tests male speaker position at 30°
4. Context dependent tests male speaker position at 60°
5. Context dependent tests male speaker position at 120°

6. Context dependent tests male speaker position at 180°

First, the audio files of female and male speakers were tested individually. These results can be found in Table 4.8. Then, we have combined all sentences of a speaker as a single audio file. Finally, we separated all combined audio files. The WERs obtained for the mixtures and separated speakers for different source positions are presented in Tables 4.3.4, 4.10, 4.11, and 4.12. *Mix_Female* and *Sep_Female* indicate that the WERs were for the female speaker from the mixture and separated signal, respectively. Similarly, *Mix_Male* and *Sep_Male* indicate that the WERs were for the male speaker output. *Mix_Ave* are the average of the results of male and female speakers.

Table4.8: Context Dependent Test Results For Individual Female and Male Speakers

Speaker	Female Speaker	Male Speaker
1	0.0214	0.0234
2	0.0286	0.0143
3	0.0143	0.0448
4	0.0214	0.0687
5	0.0143	0.0143
6	0.0143	0.0214
7	0.0143	0.0448
8	0.0291	0.0286
9	0.0214	0.0497
10	0.0143	0.0214
Average	0.01934	0.03314

Speakers	*Mix_Female*	*Mix_Male*	*Mix_Ave*	*Sep_Female*	*Sep_Male*
F1_M1	0.2936	1.0154	0.6545	**0.1079**	10.510
F2_M2	0.8327	0.2869	**0.5598**	1.7665	0.3939
F3_M3	1.0647	**0.0817**	0.5732	5.5683	**0.3274**
F4_M4	1.0798	0.1961	0.6379	0.6420	1.5517
F5_M5	0.9164	0.8864	0.9014	0.2152	2.9310
F6_M6	0.9834	0.5893	0.7863	0.3519	1.3572
F7_M7	**0.2398**	1.0486	0.6442	0.1196	8.1230
F8_M8	0.9354	0.5357	0.7355	3.7522	3.0297
F9_M9	0.7930	0.7125	0.7527	1.0628	1.7208
F10_M10	0.8982	0.9130	0.9056	0.6782	2.5537
Average	0.8037	0.6265	0.7151	1.4264	2.3039

Table4.9: Context Dependent Test Results for the Case When Male Speaker Position is at 30°

When WERs are investigated in Table 4.8, it can be seen that all WERs for both female and male speakers are very low, i.e. they have been recognized with high accuracy. The values are even lower than the average WERs specified by the Google cloud system. When we examine the WERs separately, in some cases both male and

Table4.10: Context Dependent Test Results for the Case When Male Speaker Position is at 60°

Speakers	*Mix_Female*	*Mix_Male*	*Mix_Ave*	*Sep_Female*	*Sep_Male*
F1_M1	0.2979	1.0025	0.6502	0.7637	10.650
F2_M2	0.8123	0.3320	0.5721	4.8450	**0.2367**
F3_M3	1.0738	**0.0644**	**0.5691**	7.7950	0.2722
F4_M4	1.1044	0.2173	0.6608	4.8542	1.8620
F5_M5	0.9058	0.8693	0.8875	**0.4331**	4.1903
F6_M6	0.9866	0.5740	0.7803	1.8428	4.7226
F7_M7	**0.2358**	1.0992	0.6675	0.6365	3.0594
F8_M8	0.9465	0.4508	0.6986	5.0733	0.7967
F9_M9	0.7641	0.6800	0.7220	1.8492	1.6009
F10_M10	1.1142	0.9813	1.0447	3.0658	4.7795
Average	0.8241	0.6270	0.7252	3.1158	2.2585

Table4.11: Context Dependent Test Results for the Case When Male Speaker Position is at 120°

Speakers	*Mix_Female*	*Mix_Male*	*Mix_Ave*	*Sep_Female*	*Sep_Male*
F1_M1	0.4681	0.9023	0.6852	2.2559	6.6595
F2_M2	0.8500	0.2290	**0.5395**	6.2250	0.1173
F3_M3	1.0647	**0.0745**	0.5696	10.5667	0.0985
F4_M4	1.0706	0.1055	0.5880	7.2671	0.2484
F5_M5	0.9941	0.3202	0.6571	1.3385	0.8841
F6_M6	1.0038	0.3616	0.6827	4.5303	0.5180
F7_M7	**0.3711**	1.0021	0.6866	**1.0705**	0.7555
F8_M8	0.9594	0.2391	0.5992	6.9167	**0.0974**
F9_M9	0.8911	0.4728	0.6819	5.8028	0.3026
F10_M10	0.9894	0.5261	0.7577	8.3100	0.7122
Average	0.8662	0.4233	0.6447	5.4283	1.0393

Table4.12: Context Dependent Test Results for the Case When Male Speaker Position is at 180°

Speakers	*Mix_Female*	*Mix_Male*	*Mix_Ave*	*Sep_Female*	*Sep_Male*
F1_M1	0.3965	1.0010	0.6987	**0.4619**	No Output
F2_M2	0.8450	0.2356	**0.5403**	5.3417	No Output
F3_M3	1.0647	**0.0817**	0.5745	11.0667	No Output
F4_M4	1.0860	0.1472	0.6166	5.1547	No Output
F5_M5	0.8520	0.5346	0.6933	0.8589	No Output
F6_M6	1.0113	0.4954	0.7533	3.3462	No Output
F7_M7	**0.2816**	0.9953	0.6384	0.5447	No Output
F8_M8	0.9688	0.3003	0.6345	5.6267	No Output
F9_M9	0.8042	0.5259	0.6650	4.1549	No Output
F10_M10	0.9012	0.6690	0.7851	4.1179	No Output
Average	0.8211	0.4986	0.6599	4.0674	NA

female WERs (1.43%) are almost one-fourth of WERs (4.9%) of Google cloud system explained. This is an unexpected situation. This situation may be explained by the following reasons;

1. Audio files were recorded in a noiseless environment.

2. Audio files are context dependent which means that the cloud system can correctly identify the sentences with predictions, even if they are misinterpreted.

3. The data sets may have been used by others before. In this case, the Google cloud system may have already been trained using these data sets.

We believe that the explanations 2 and 3 mentioned above are very likely, because we have used prerecorded audio files which anyone can access through the use of the internet and Google cloud system is known to exploit the context to make a prediction [81].

In Table 4.3.4, the mixtures and separated audio files were examined for the case when the female speaker is positioned at 0°, and the male speaker is positioned at 30°. The following results were observed:

- The lowest WER is achieved for the female speaker output of the mixture when speakers are F7_M7 and, the highest WER is achieved when speakers are F4_M4.

- The lowest WER is achieved for the male speaker output of the mixture when speakers are F3_M3 and, the highest WER is achieved when speakers are F7_M7.

- The average of WERs when the output is calculated according to the female speaker output of the mixtures is 80.37% and according to the male speaker output is 62.65%, which are close to each other.

- The lowest WER is achieved for the female speaker output of the separated sounds when speakers are F1_M1 and the highest WER is achieved when speakers are F3_M3.

- The lowest WER is achieved for the male speaker output of the separated sounds when speakers are F3_M3 and the highest WER is achieved when speakers are F1_M1.

- The average of WERs when the output is calculated according to the female speaker output of the separated sounds is 67.82% and according to the male speaker output is 255.37%.

In Table 4.10, the mixtures and separated audio files were examined for the case when the female speaker is positioned at 0°, and the male speaker is positioned at 60°. The following results were observed:

- The lowest WER is achieved for the female speaker output of the mixture when speakers are F7_M7 and, the highest WER is achieved when speakers are F4_M4.
- The lowest WER is achieved for the male speaker output of the mixture when speakers are F4_M4 and, the highest WER is achieved when speakers are F7_M7.
- The average of WERs when the output is calculated according to the female speaker output of the mixtures is 111.42% and according to the male speaker output is 98.13%.
- The lowest WER is achieved for the female speaker output of the separated sounds when speakers are F5_M5 and the highest WER is achieved when speakers are F3_M3.
- The lowest WER is achieved for the male speaker output of the separated sounds when speakers are F8_M8 and the highest WER is achieved when speakers are F1_M1.
- The average of WERs when the output is calculated according to the female speaker output of the separated sounds is 306.58% and according to the male speaker output is 477.95%.

In Table 4.11, the mixtures and separated audio files were examined for the case when the female speaker is positioned at 0°, and the male speaker is positioned at 120°. The following results were observed:

- The lowest WER is achieved for the female speaker output of the mixture when speakers are F7_M7 and, the highest WER is achieved when speakers are F4_M4.
- The lowest WER is achieved for the male speaker output of the mixture when speakers are F4_M4 and, the highest WER is achieved when speakers are F7_M7.
- The average of WERs when the output is calculated according to the female speaker output of the mixtures is 98.94% and according to the male speaker output is 52.61%.
- The lowest WER is achieved for the female speaker output of the separated sounds when speakers are F7_M7 and the highest WER is achieved when speakers are F3_M3.
- The lowest WER is achieved for the male speaker output of the separated sounds when speakers are F8_M8 and the highest WER is achieved when speakers are F1_M1.
- The average of WERs when the output is calculated according to the female speaker output of the separated sounds is 831.00% and according to the male speaker output is 71.22%.

In Table 4.12, the mixtures and separated audio files were examined for the case when the female speaker is positioned at 0°, and the male speaker is positioned at 180°. The following results were observed:

- The lowest WER is achieved for the female speaker output of the mixture when speakers are F7_M7 and, the highest WER is achieved when speakers are F4_M4
- The lowest WER is achieved for the male speaker output of the mixture when speakers are F3_M3 and, the highest WER is achieved when speakers are F1_M1.
- The average of WERs when the output is calculated according to the female speaker output of the mixtures is 90.12% and according to the male speaker output is 66.90%.
- The lowest WER is achieved for the female speaker output of the separated sounds when speakers are F1_M1 and the highest WER is achieved when speakers are F3_M3.
- No output was produced in case of male speaker output of the separated sounds.
- The average of WERs when the output is calculated according to the female speaker output of the separated sounds is 411.79%.

After careful consideration of all observations, it can be said that there is no single mixture or separated output that always provides the lowest WER or the highest WER. Although there are exceptions, in general if an output (separated or mixture) gives the highest (or lowest) WER for one speaker, it gives the lowest (or highest) WER for the other speaker. This means one of the speakers usually dominates the mixture and is separated better.

Generally speaking, it is expected that the WER for separated sounds should be lower than the WER of corresponding mixture. However, the results show that it is not always true. For example, in Table 4.3.4, for the speakers F7_M7, the WER calculated for male speaker output for the mixture was 104.86% and for separated output 812.30%. There are also examples of the opposite. It can be explained by the calculation of WER. When mixtures are recognized, the number of deletions would usually be 0, as there are more spoken words in the test sample than in the reference. Therefore, the misrecognized words are considered as substitutions. Since in the mixtures, the speeches of both speakers overlap, the number of insertions would also be limited. This artificially lowers the WERs for the mixtures. In fact, manual comparison of the recognized sentences for the mixtures and the separated sounds reveals that the recognized sentences for the separated sounds are more correct and meaningful than the recognized sentences for the mixtures considering their content, although the WER is higher for the former than the latter.

In some cases, the WER is too high. It is even larger than 500%. When we examined the outputs for these cases, it was seen that the cloud system did not produce an output or a sentence that consisted of several words; instead, just one or two words were output. For example, in Table 4.3.4, for mixture of the speakers F1_M1 the output

was *Philosophers Education*, for which the correct output should be *Philosophers of education often differ in their views on the nature of knowledge*. Since the number of deletions in this case is very high, this leads to a very high WER. It is thought that such errors might be due to a problem in the real-time transfer of speech frames in between mobile application and cloud system, such as due to packet loss, rather than a problem in the recognition itself.

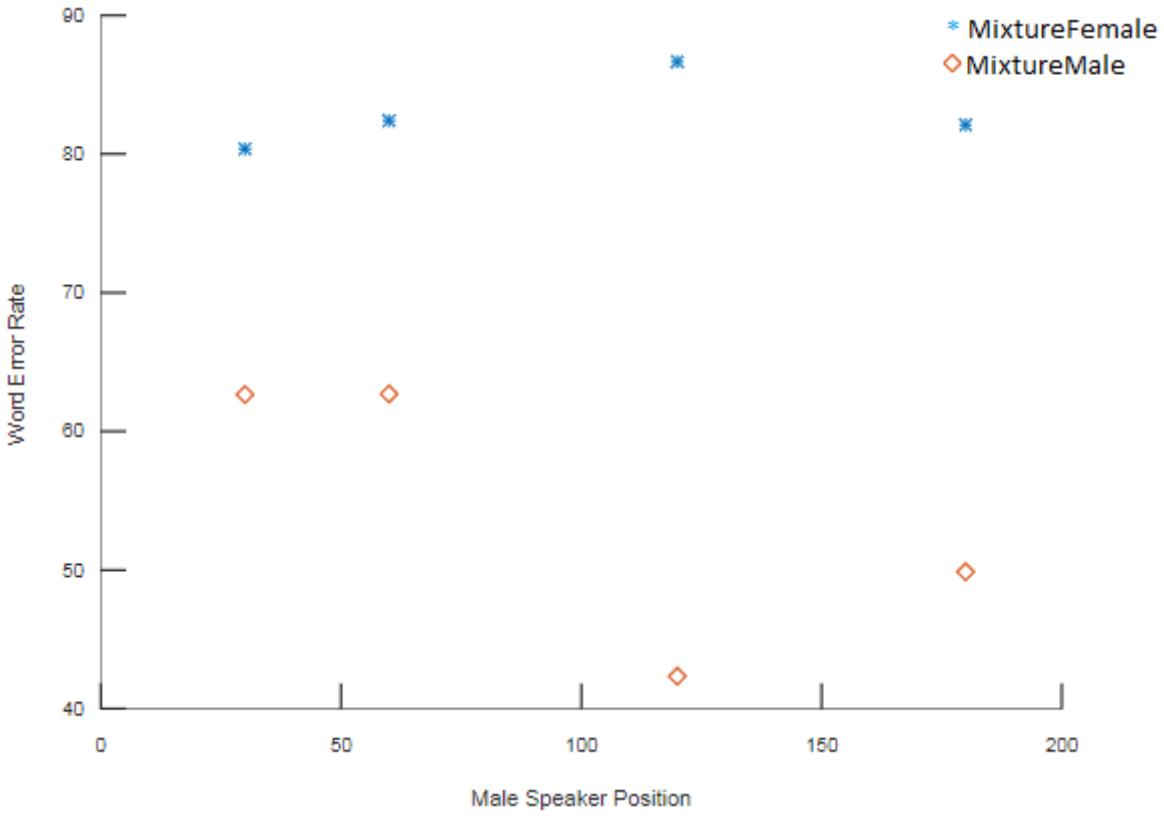


Figure 4.6: Context Dependent Mixture Test Different Male Speaker Position Results

As it can be seen from Figure 4.6,the WERs for the mixtures for changing source direction is almost the same for the female speech, but different for the male speech. It is due to the fact that for these test cases the direction of the female speaker stays the same at 0°, but the direction of the male speaker changes. Although the distances between the speakers and the microphone array are the same in all cases, the direction change in the male speaker's position may lead to higher or lower signal levels due to constructive or destructive additions of early reflections. Therefore, the signal levels of male speech would be different for different directions. The lowest WER is achieved when male speaker is positioned at 120°.

As it can be seen from Figure 4.7, the average of WERs calculated according to the male speech outputs is lower than the female speech outputs. The WERs calculated according to the separated female speech output increases with increasing angular difference between the speakers, since the separation performance also increases. However, at 180°, there seems to be a problem in the separation performance. For this direction, there is no output produced for the male speech.

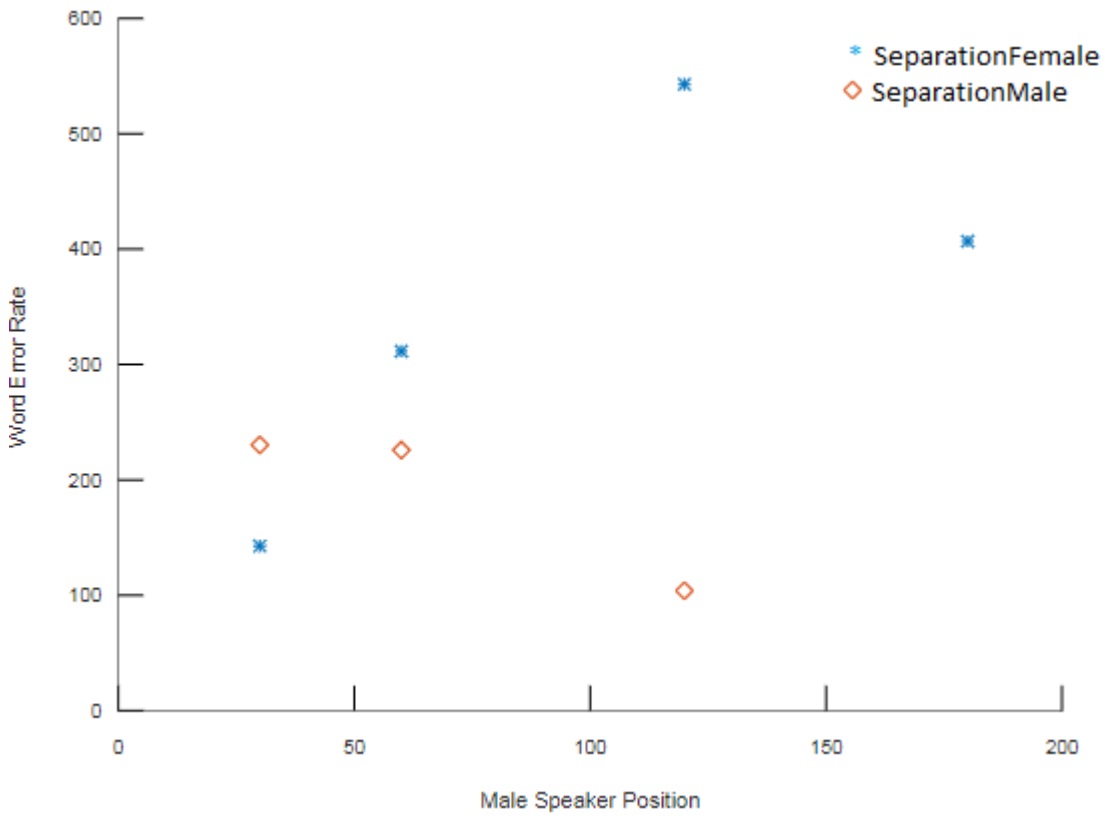


Figure 4.7: Context Dependent Separation Test Male Speaker Position Results

CHAPTER 5

DISCUSSIONS

The results of all the performed tests are given in the previous section. Here, we will interpret the results and compare them with the studies in the literature.

By looking at the WERs, we have compared the speech recognition performances of different cloud systems. Google cloud has clearly produced better speech recognition results than other systems. The study also showed that Google achieved 1.50 times better performance than Microsoft and 1.81 times better performance than IBM [82]. It also produced output for noisy recordings, which shows that it is more robust against noise than other systems. Our findings are in line with other studies which compared these systems. For example, in a study conducted in 2017 [83], Google's recognition performance was found to be twice better than Microsoft's speech recognition performance, with WERs of 9% by Google and 18% by Microsoft while using clean audio files. In another study performed in 2017 [84], Google achieved 1.3 times better performance than IBM and 1.6 times better performance than Microsoft. In our study, Google achieved 5.43% while Microsoft achieved 44.35%. When we checked the studies that compare recognition performances, it was found that different tests have been performed as well as different metrics, most of which were mentioned in Section 4.1. However, Google was found to be superior in most cases.

The gender bias in Google cloud's speech recognition performance has been studied before [85], which found that male speech is recognized much better than female speech. Google also released an application that makes speech recognition for children [86] and in this study, WER for female speech was found to be lower than for male speech. In our study, we have also found that for the mixtures of two speakers, the male speech is recognized with higher accuracy than female speech. However, it was vice versa when they were tested individually. Also, in the mixtures, the signal levels of both speakers are not always equal, which might explain discrepancies in the performance. Therefore, our results do not indicate gender bias in the recognition performance. When the speech signals were contaminated with restaurant background noise, the average WER was found to be 0.822. After noise cancellation, the average WER reduced to 0.092. Accuracy is improved 5.1 times. This improvement supported our expectations from the system.

As can be seen in Figures 4.4, and 4.5, speech recognition performance is negatively affected by noise. When the noise level increases, speech intelligibility decreases and this leads to higher WERs. However, when noise cancellation is applied, the WER stays almost constant and is not affected much by noise. This also realizes our expectations

from the system.

As it can be inferred from Tables 4.4 and 4.1 that provide results for context independent tests and context independent rhyme tests, respectively, better accuracy is achieved for the former, i.e., when words are used in a sentence, rather than individually. While the context independent tests were in the form of *"please select the word went"* where only the last word changes, in context independent rhyme tests only one word was used. Misrecognition of the last word in the former test would result in a WER of 0.2 (1 substitution/5 words), whereas in the latter test 1 (1 substitution/1 word). Despite this bias, using words in a sentence results in only 1.25 times better recognition performance.

In Table 4.3.4, 4.10, 4.11, and 4.12, the mixtures and the separated audio files were examined according different positioning of the speakers, resulting in different angular spacing between them, thereby affecting the separation performance. Our expectations were the following;

1. WERs for clean, i.e., interference free speech signals would be lower than the WERs obtained for mixtures and separated signals
2. WERs for separated signals would be lower than the WERs obtained for mixtures.
3. With increasing angular separation between the speakers, the WERs for separated signals would decrease.

Except for expectation 1, others were not fully realized. Separated signals resulted in higher WERs than the mixtures. In such cases, intersection of sets of words obtained from the reference and the test sample could be used for better speech recognition performance evaluation. While the WERs for separated male speech decreases (with the exception of the 180°, where no output was received), the WERs for separated female speech increases with increasing angular spacing. The reason for these could be explained by the difference in the number of words present in mixtures and in separated signals. Since, there are fewer words in the latter, WER calculation penalizes the separated signals.

CHAPTER 6

CONCLUSIONS

In this thesis study, a portable real-time speech recognition system for mobile devices was proposed with the aim of increasing the speech recognition performance in noisy environments. Main limiting design constraints were real-time operation, a small size for the microphone array, capturing signals from all directions without any preferred direction of operation and limited processing power of mobile devices. Therefore, the design handled the noise cancellation operation in a separate accessory which transferred the denoised speech to the mobile device. The mobile device was then used to forward this signal to the cloud and display the received response. Running the noise cancellation algorithms on a mobile device was also considered. However, this option has not been selected due to the resource limitation problem, found in most battery powered devices.

The designed system is different from the existing solutions, which cannot provide sufficient speech recognition accuracy in noisy environments. Instead of conventional noise cancellation algorithms, which depend on the low-frequency noise assumption, a system specific algorithm was utilized in the designed system, which is based on source separation using a compact microphone array.

The designed system is a cloud based system. We used different cloud system speech APIs provided by IBM, Microsoft, and Google, to compare their recognition rates (see Section 4.3.1). After making careful observations, it was revealed that Google performs better than other cloud systems. While IBM was slightly better than Microsoft, Google was the only cloud system that provided an output for noisy recordings. Google's speech recognition is context dependent and works quite successfully in a noiseless environment. For example, a sentence like "What is the weather life" would automatically be corrected as "What is the weather like". Human beings also perceive speech according to the context and make such corrections. When speech is present in the form of a meaningful sentence, rather than a collection of words, the recognition gets better.

We utilized different audio datasets that included context independent data tests, context independent rhyme tests, context independent tests with different SNR test and context dependent test in order to show the effect of the noise, the noise level, the speaker and the gender of the speaker on speech recognition performance. The outcomes of the tests showed that the Google cloud's speech recognition performance does depend on the speaker or the gender of the speaker.

The detailed results show that WERs are too high for noisy audio files. This indicates that applications based on speech recognition with mobile devices cannot be used in noisy environments. However, the speech recognition performance obtained after applying a noise reduction algorithm through the designed system was very close to that of the clean recordings.

The performance of the designed system in noisy environments is well (see Section 4.3.3) and thus the system is not affected until the SNR is equal to 9 dB. Moreover, running the developed application on mobile devices improves the usability of the system and makes it portable. And also, using a tetrahedral microphone array enables the designed system to capture sounds from all directions. Besides all the features of the designed system, working in real-time was the most challenging feature.

In the future, we would like to test the performance of the system under wind noise and if necessary develop a specific wind noise removal algorithm. It is also among the plans to make the performance measurements for the Turkish language and compare them with the English language.

REFERENCES

[1] K. S. Helfer and R. L. Freyman, "Aging and speech-on-speech masking," *Ear and hearing*, vol. 29, no. 1, p. 8, 2008.

[2] J. Y. Lee, J. T. Lee, H. J. Heo, C.-H. Choi, S. H. Choi, and K. Lee, "Speech recognition in real-life background noise by young and middle-aged adults with normal hearing," *Journal of audiology & otology*, vol. 19, no. 1, p. 39, 2015.

[3] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Communications of the ACM*, vol. 57, no. 1, pp. 94–103, 2014.

[4] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system : A comparative study," *CoRR*, vol. abs/1305.1145, 2013.

[5] S. Saksamudre, P. Shrishrimal, and R. Deshmukh, "A review on different approaches for speech recognition system," vol. 115, pp. 23–28, 04 2015.

[6] S. Narang and M. D. Gupta, "Speech feature extraction techniques : A review," 2015.

[7] N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition," *International Journal For Advance Research in Engineering And Technology(ISSN 2320-6802)*, vol. Volume 1, 07 2013.

[8] A. Oirere, G. Janvale, and R. R, "Automatic speech recognition and verification using lpc, mfcc and svm," *International Journal of Computer Applications*, vol. 127, pp. 47–52, 10 2015.

[9] N. Singh, R. A. Khan, and R. Shree, "Article: Mfcc and prosodic feature extraction techniques: A comparative study," *International Journal of Computer Applications*, vol. 54, pp. 9–13, September 2012.

[10] "Acoustic modeling." `https://www.microsoft.com/en-us/research/project/acoustic-modeling/`. [Online; accessed 10-April-2018].

[11] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb 1989.

[12] M. Vyas, "A gaussian mixture model based speech recognition system using matlab," *Signal Image Processing : An International Journal*, vol. 4, pp. 109–118, 08 2013.

[13] S. S, P. L, and S. S, "A review on automatic speech recognition system in indian regional languages," vol. 181, pp. 38–42, 07 2018.

[14] K. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[15] "A brief history of voice recognition technology." https://www.totalvoicetech.com/a-brief-history-of-voice-recognition-technology.html, Jul 2016. [Online; accessed 07-April-2018].

[16] D. H. Klatt, "Review of arpa speech understanding project," *Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1345–1366, 1977.

[17] B. Lowerre and R. Reddy, "The harpy speech understanding system," in *Readings in speech recognition*, pp. 576–586, Elsevier, 1990.

[18] X. Huang, F. Alleva, M.-Y. Hwang, and R. Rosenfeld, "An overview of the sphinx-ii speech recognition system," in *Proceedings of the workshop on Human Language Technology*, pp. 81–86, Association for Computational Linguistics, 1993.

[19] J. Markoff, "Scientists see advances in deep learning, a part of artificial intelligence." https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html?pagewanted=1&_r=3, Nov 2012. [Online; accessed 15-April-2018].

[20] L. R. Rabiner, "Applications of voice processing to telecommunications," *Proceedings of the IEEE*, vol. 82, no. 2, pp. 199–228, 1994.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[22] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černockỳ, "Strategies for training large scale neural network language models," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pp. 196–201, IEEE, 2011.

[23] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*, pp. 8614–8618, IEEE, 2013.

[24] B. Sosinsky, "Cloud computing," 2011.

[25] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?," *Computer*, vol. 43, no. 4, pp. 51–56, 2010.

[26] I. A. Elgendy, M. El-kawkagy, and A. Keshk, "Improving the performance of mobile applications using cloud computing," in *Informatics and Systems (INFOS), 2014 9th International Conference on*, pp. PDC–109, IEEE, 2014.

[27] J. Kincaid, "Google privacy blunder shares your docs without permission," *TechCrunch, March*, 2009.

[28] R. McMillan, "Hacker: I broke into twitter," *Network World, available at: http://tinyurl. com/cdmzqc (www. networkworld. com/news/2009/043009-hacker-i-broke-into. html*, 2009.

[29] M. Schuster, "Speech recognition for mobile devices at google," in *Pacific Rim International Conference on Artificial Intelligence*, pp. 8–10, Springer, 2010.

[30] J. Novet, "Google says its speech recognition technology now has only an 8word error rate." https://venturebeat.com/2015/05/28/google-says-its-speech-recognition-technology-now-has-only-an-8-word-error-rate May 2015. [Online; accessed 07-March-2018].

[31] "Speech api - speech recognition | google cloud." https://cloud.google.com/speech/. [Online; accessed 01-November-2017].

[32] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[33] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, "The ibm 2015 english conversational telephone speech recognition system," *arXiv preprint arXiv:1505.05899*, 2015.

[34] "Ibm cloud." https://console.bluemix.net/docs/services/speech-to-text/index.html#about. [Online; accessed 21-March-2017].

[35] "Msdn magazine." https://web.archive.org/web/20080307054756/http://msdn2.microsoft.com/en-us/magazine/cc163663.aspx. [Online; accessed 08-April-2018].

[36] A. Linn, "Speak, hear, talk: The long quest for technology that understands speech as well as a human." https://news.microsoft.com/features/speak-hear-talk-the-long-quest-for-technology-that-understands-speech-as-well-a#sm.0000jro6m9fcgfhtuhu2e1z6uoaa9, Dec 2015. [Online; accessed 18-March-2018].

[37] R. Eckel, "Microsoft researchers achieve speech recognition milestone." https://blogs.microsoft.com/ai/microsoft-researchers-achieve-speech-recognition-milestone/#sm.0000jro6m9fcgfhtuhu2e1z6uoaa9, Sep 2016. [Online; accessed 08-March-2018].

[38] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The microsoft 2016 conversational speech recognition system," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 5255–5259, IEEE, 2017.

[39] "Number of smartphone users worldwide 2014-2020." https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/. [Online; accessed 23-March-2018].

[40] M. Meeker, "Internet trends 2017. code conference. 31 may 2017," 2017.

[41] D. O'Shaughnessy, "Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[42] X. Huang and K.-F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Transactions on Speech and Audio processing*, vol. 1, no. 2, pp. 150–157, 1993.

[43] M. Assefi, M. P. Wittie, and A. Knight, "Impact of network performance on cloud speech recognition. icccn," *IEEE. Aug*, 2015.

[44] G. A. Flamme, M. R. Stephenson, K. Deiters, A. Tatro, D. Van Gessel, K. Geda, K. Wyllys, and K. McGregor, "Typical noise exposure in daily life," *International journal of audiology*, vol. 51, no. sup1, pp. S3–S11, 2012.

[45] L. E. Humes, "Speech understanding in the elderly," *Journal-American Academy of Audiology*, vol. 7, pp. 161–167, 1996.

[46] C. E. Johnson, "Childrens' phoneme identification in reverberation and noise," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 1, pp. 144–157, 2000.

[47] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[48] M. A. Bee and C. Micheyl, "The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it?," *Journal of comparative psychology*, vol. 122, no. 3, p. 235, 2008.

[49] A. Koutras and E. Dermatas, "Robust speech recognition in a high interference real room environment using blind speech extraction," in *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, vol. 1, pp. 167–171, IEEE, 2002.

[50] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.

[51] A. C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein, "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear and hearing*, vol. 31, no. 3, pp. 336–344, 2010.

[52] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[53] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574 – 584, 2015.

[54] N. Sharma and S. Sardana, "A real time speech to text conversion system using bidirectional kalman filter in matlab," pp. 2353–2357, 09 2016.

[55] Y. H. Goh, P. Raveendran, and Y. L. Goh, "Robust speech recognition system using bidirectional kalman filter," *IET Signal Processing*, vol. 9, no. 6, pp. 491–497, 2015.

[56] J. Jebastine and S. Rani, "Design and implementation of noise free audio speech signal using fast block least mean square algorithm," vol. 3, 06 2012.

[57] M. A. J. Sathya and D. S. P. Victor, "Noise reduction techniques and algorithms for speech signal processing," 2015.

[58] H. Adel, M. Souad, A. Alaqeeli, and A. Hamid, "Beamforming techniques for multichannel audio signal separation," *arXiv preprint arXiv:1212.6080*, 2012.

[59] D. H. Johnson and D. E. Dudgeon, "Array signal processing," 1993.

[60] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 489–498, Sept 2004.

[61] A. K. Kattepur, J. P. Lie, F. Sattar, and C. M. S. See, "High fidelity blind source separation of speech signals," in *2009 17th European Signal Processing Conference*, pp. 859–863, Aug 2009.

[62] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512)*, vol. 5, pp. V–V, May 2004.

[63] V. G. Reju, S. N. Koh, I. Y. Soon, and X. Zhang, "Solving permutation problem in blind source separation of speech signals: A method applicable for collinear sources," in *2005 5th International Conference on Information Communications Signal Processing*, pp. 1461–1465, Dec 2005.

[64] R. Acharyya and J. Flierl, "Blind source separation by time frequency masking of an underdetermined system," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pp. 2228–2231, May 2008.

[65] U. Kjems, M. Pedersen, J. Boldt, T. Lunner, and D. Wang, "Speech intelligibility of ideal binary masked mixtures," 09 2018.

[66] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[67] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[68] E. Ferro and F. Potorti, "Bluetooth and wi-fi wireless protocols: a survey and a comparison," *IEEE Wireless Communications*, vol. 12, no. 1, pp. 12–26, 2005.

[69] J.-S. Lee, Y.-W. Su, and C.-C. Shen, "A comparative study of wireless protocols: Bluetooth, uwb, zigbee, and wi-fi," in *Industrial Electronics Society, 2007. IECON 2007. 33rd Annual Conference of the IEEE*, pp. 46–51, Ieee, 2007.

[70] L. Deng and X. Huang, "Challenges in adopting speech recognition," *Communications of the ACM*, vol. 47, no. 1, pp. 69–75, 2004.

[71] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 748–756, 2008.

[72] K. Ozeki and N. Hamada, "Estimating directions of multiple sound sources using tetrahedral microphone array," in *TENCON 2006. 2006 IEEE Region 10 Conference*, pp. 1–4, IEEE, 2006.

[73] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Intensity vector direction exploitation for exhaustive blind source separation of convolutive mixtures," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 41–44, IEEE, 2009.

[74] B. Günel, "On the statistical distributions of active intensity directions," *Journal of Sound and Vibration*, vol. 332, no. 20, pp. 5207–5216, 2013.

[75] J. Vincent, "99.6 percent of new smartphones run android or ios." `https://www.theverge.com/2017/2/16/14634656/android-ios-market-share-blackberry-2016`, Feb 2017. [Online; accessed 21-March-2018].

[76] "Mobile operating system market share worldwide." `http://gs.statcounter.com/os-market-share/mobile/worldwide`. [Online; accessed 08-April-2018].

[77] S. Voran, "Using articulation index band correlations to objectively estimate speech intelligibility consistent with the modified rhyme test," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pp. 1–4, IEEE, 2013.

[78] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *The Journal of the Acoustical Society of America*, vol. 30, no. 7, pp. 596–600, 1958.

[79] A. S. House, C. E. Williams, M. H. Hecker, and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *The Journal of the Acoustical Society of America*, vol. 37, no. 1, pp. 158–166, 1965.

[80] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.

[81] "Speech api - speech recognition | google cloud." [Online; accessed 01-April-2018].

[82] J. Kincaid, "Which automatic transcription service is the most accurate?—2018," Sep 2018. [Online; accessed 13-February-2019].

[83] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (microsoft api, google api and cmu sphinx)," *Int. J. Eng. Res. Appl*, vol. 7, pp. 20–24, 2017.

[84] O. Biran, "You shall not speak: Benchmarking speech recognition apis for bots." `https://recast.ai/blog/benchmarking-speech-recognition-api/`, Dec 2017. [Online; accessed 22-April-2018].

[85] S. Larson, "Research shows gender bias in google's voice recognition." `https://www.dailydot.com/debug/google-voice-recognition-gender-bias/`, Jul 2016. [Online; accessed 05-April-2018].

[86] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015.

APPENDIX A

LIST OF WORD GROUPS USED IN CONTEXT INDEPENDENT RHYME TESTS

1. Word Group: went, sent, bent, dent, tent, rent
2. Word Group: hold, cold, told, fold, sold, gold
3. Word Group: pat, pad, pan, path, pack, pass
4. Word Group: lane, lay, late, lake, lace, lame
5. Word Group: kit, bit, fit, hit, wit, sit
6. Word Group: must, bust, gust, rust, dust, just
7. Word Group: teak, team, teal, teach, tear, tease
8. Word Group: din, dill, dim, dig, dip, did
9. Word Group: bed, led, fed, red, wed, shed
10. Word Group: pin, sin, tin, fin, din, win
11. Word Group: dug, dung, duck, dud, dub, dun
12. Word Group: sum, sun, sung, sup, sub, sud
13. Word Group: seep, seen, seethe, seek, seem, seed
14. Word Group: not, tot, got, pot, hot, lot
15. Word Group: vest, test, rest, best, west, nest
16. Word Group: pig, pill, pin, pip, pit, pick
17. Word Group: back, bath, bad, bass, bat, ban
18. Word Group: way, may, say, pay, day, gay
19. Word Group: pig, big, dig, wig, rig, fig
20. Word Group: pale, pace, page, pane, pay, pave
21. Word Group: cane, case, cape, cake, came, cave
22. Word Group: shop, mop, cop, top, hop, pop
23. Word Group: coil, oil, soil, toil, boil, foil
24. Word Group: tan, tang, tap, tack, tam, tab
25. Word Group: fit, fib, fizz, fill, fig, fin
26. Word Group: same, name, game, tame, came, fame
27. Word Group: peel, reel, feel, eel, keel, heel
28. Word Group: hark, dark, mark, bark, park, lark
29. Word Group: heave, hear, heat, heal, heap, heath
30. Word Group: cup, cut, cud, cuff, cuss, cub
31. Word Group: thaw, law, raw, paw, jaw, saw
32. Word Group: pen, hen, men, then, den, ten
33. Word Group: puff, puck, pub, pus, pup, pun
34. Word Group: bean, beach, beat, beak, bead, beam
35. Word Group: heat, neat, feat, seat, meat, beat

36. Word Group: dip, sip, hip, tip, lip, rip
37. Word Group: kill, kin, kit, kick, king, kid
38. Word Group: hang, sang, bang, rang, fang, gang
39. Word Group: took, cook, look, hook, shook, book
40. Word Group: mass, math, map, mat, man, mad
41. Word Group: ray, raze, rate, rave, rake, race
42. Word Group: save, same, sale, sane, sake, safe
43. Word Group: fill, kill, will, hill, till, bill
44. Word Group: sill, sick, sip, sing, sit, sin
45. Word Group: bale, gale, sale, tale, pale, male
46. Word Group: wick, sick, kick, lick, pick, tick
47. Word Group: peace, peas, peak, peach, peat, peal
48. Word Group: bun, bus, but, bug, buck, buff
49. Word Group: sag, sat, sass, sack, sad, sap
50. Word Group: fun, sun, bun, gun, run, nun

APPENDIX B

LIST OF SENTENCES USED IN CONTEXT DEPENDENT TESTS

1. The female produces a litter of two to four young in November and December.

2. Their solution requires development of the human capacity for social interest.

3. His most significant scientific publications were studies of birds and animals.

4. In recent years she has primarily appeared in television films such as Little Gloria.

5. Unusually high levels of radiation were detected in many European countries.

6. For the first time in years the Republicans also captured both houses of Congress.

7. The South Carolina educational radio network has won national broadcasting awards.

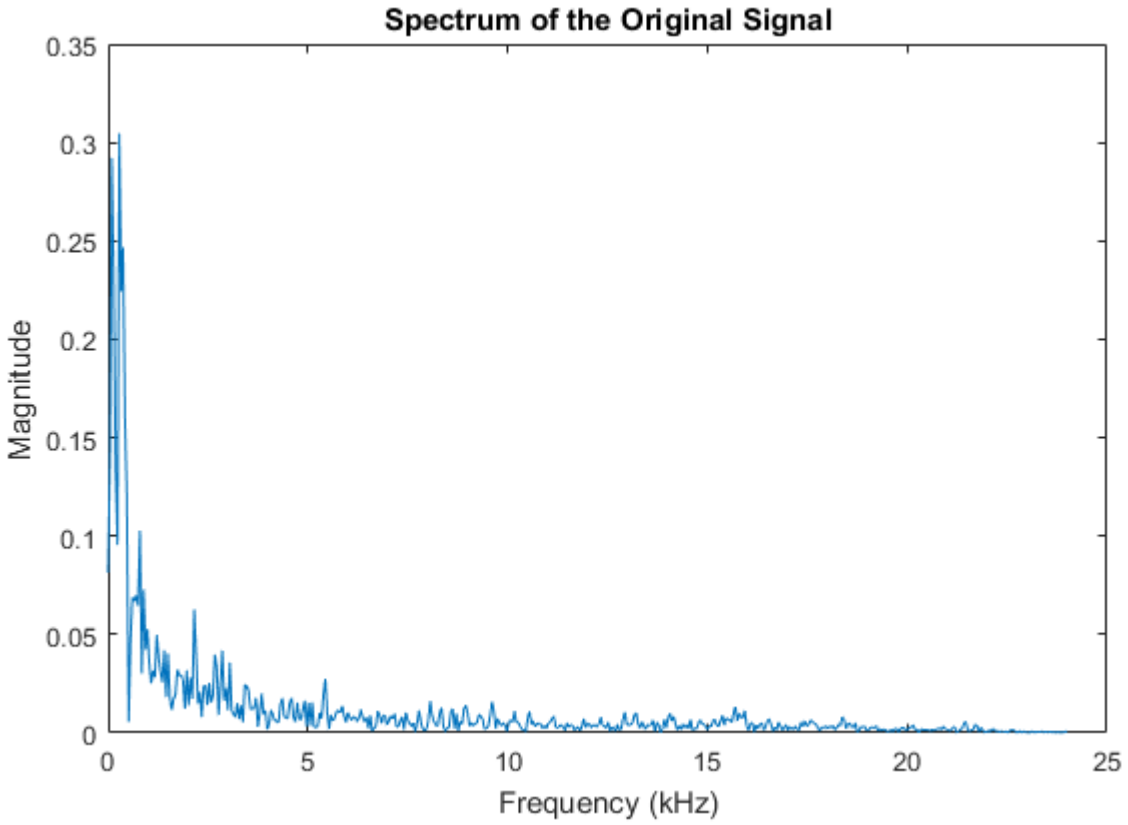
8. Modern electronics has become highly dependent on inorganic chemistry.

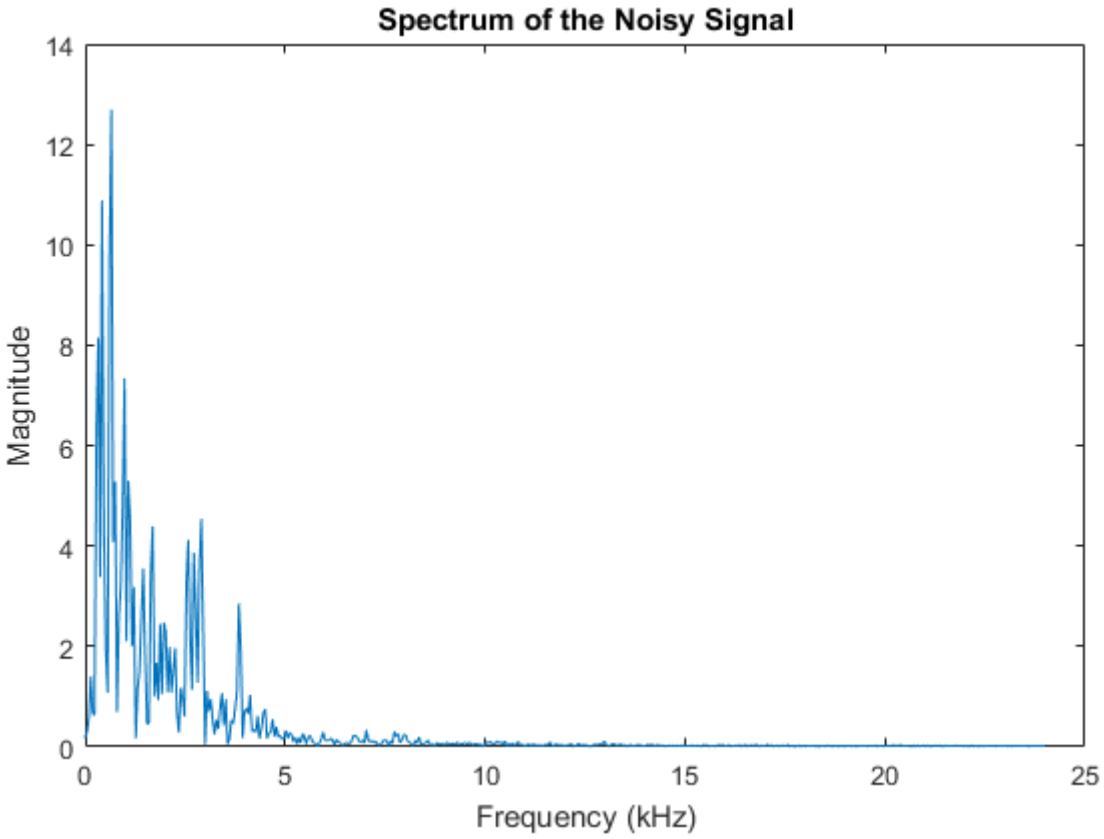
9. Much of the ground beef consumed in the United States comes from Derrick House

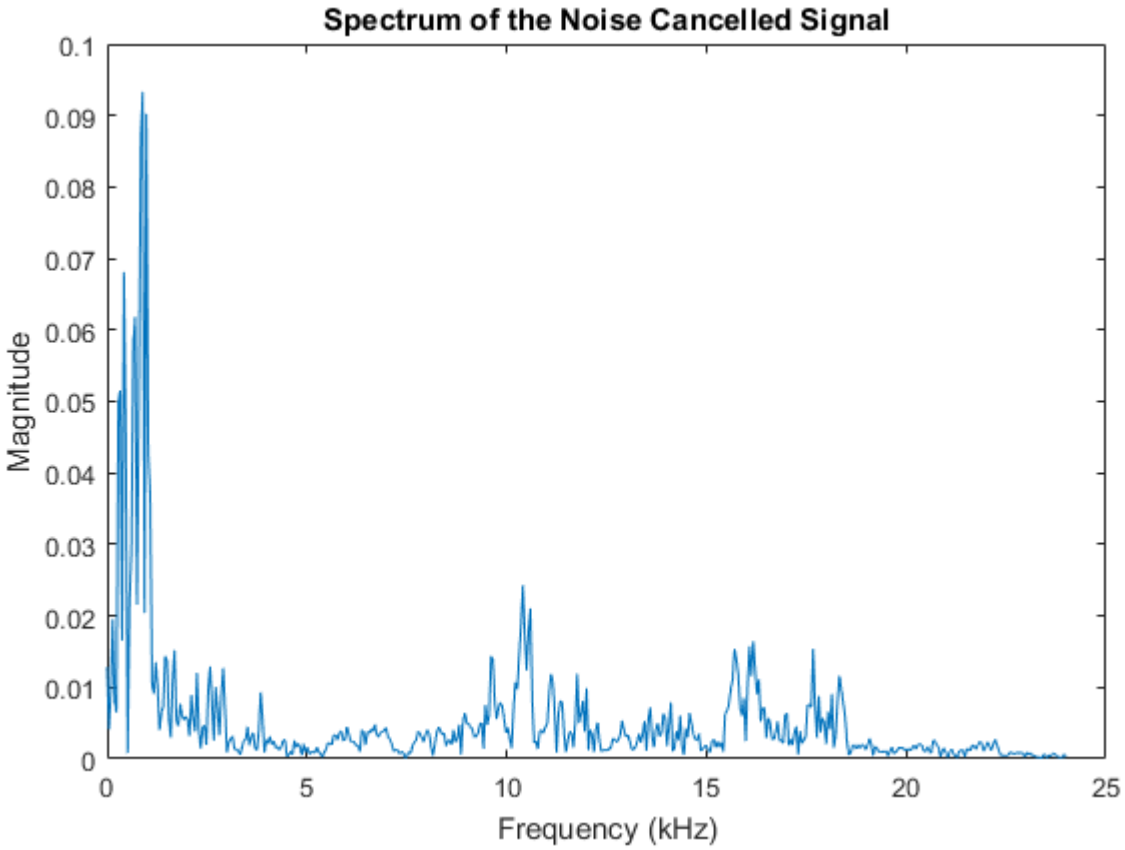
10. Philosophers of education often differ in their views on the nature of knowledge

APPENDIX C

SPECTRUM OF THE ORIGINAL, NOISY, AND NOISE CANCELLED SIGNALS WITH 3 DB SNR




Spectrum of the Noisy Signal
Magnitude
Frequency (kHz)
0
2
4
6
8
10
12
14
0
5
10
15
20
25


Spectrum of the Noise Cancelled Signal
Magnitude
Frequency (kHz)
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
5
10
15
20
25