WORKING TITLE: AN APPROACH TO AUTOMATIC AND HUMAN SPEECH RECOGNITION USING EAR-RECORDED SPEECH.

by

Samuel John Charles Johnston

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2017

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Samuel John Charles Johnston
entitled Working Title: An approach to automatic and human speech recognition using ear-recorded speech.
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 7 August 2017

Mike Hammond

Date: 7 August 2017

Brad Story

Date: 7 August 2017

Natasha Warner

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 7 August 2017

Dissertation Director: Mike Hammond

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Samuel John Charles Johnston

ACKNOWLEDGEMENTS

Insert your acknowledgements here.

This should be one page maximum, and is single-spaced by default.

DEDICATION

Insert your dedication here

One page maximum.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This is where the body of your abstract goes, limited to 150 words for a thesis, and 350 words for a dissertation or document. The word count limits apply to the regular Abstract in the thesis and to the separate Special Abstract. Use the same text for both; just adjust the margins and heading. The abstract should summarize your work. The UMI booklet listed in the resources section of the U of A manual provides some writing tips. The abstract for a dissertation or document may be longer than one page; word count is more important than page length in this section.

If you are doing a paper submission, submit one copy of the special abstract, and two extra copies of your title page, in the box with the final copies of your thesis. If you are doing an electronic submission, you can ignore the special abstract.

[dissertation,copyright]uathesis []graphicx[]color

framed

alltt

booktabs graphicx natbib

wrapfig

caption subcaption tipa color,soul url blindtext [inline]enumitem breakurl mathtools amsmath

[driverfallback=dvips,bookmarks,colorlinks=true,urlcolor=black,linkcolor=black,citecolor=blac

CHAPTER 1

Introduction

Speech in a noisy background presents a challenge to the recognition of that speech both by human listeners and by computers tasked with recognizing human speech (termed automatic speech recognition). Years of research have resulted in many solutions, though none so far have completely solved the problem. Most approaches have taken the route of trying to remove noise from a signal that is already corrupted with noise.

This project presents a new approach to the problem. It proposes that human speech can be recorded in a manner that largely eliminates the noise before it reaches the microphone recording the signal. The microphone would be placed inside the ear canal of the speaker and would record speech as it is passed through the bone and tissue of the speaker's head. The signal is expected to be distorted by passage through the head, but this distortion is hypothesized to be at least partially reversible.

Section 1.1 below describes the basic acoustics of sound, leading up to a brief overview of sound in noise. Chapter 2**??** discusses the collection of ear-recorded speech, Chapter 3**??** describes a human speech perception experiment using ear-recorded speech, and Chapter 4**??** outlines an experiment testing the ability of an automatic speech recognition system to accurately recognize ear-recorded speech. At the end of this chapter, Section 1.2 will give a more detailed overview of the rest of the dissertation.

1.1 Background

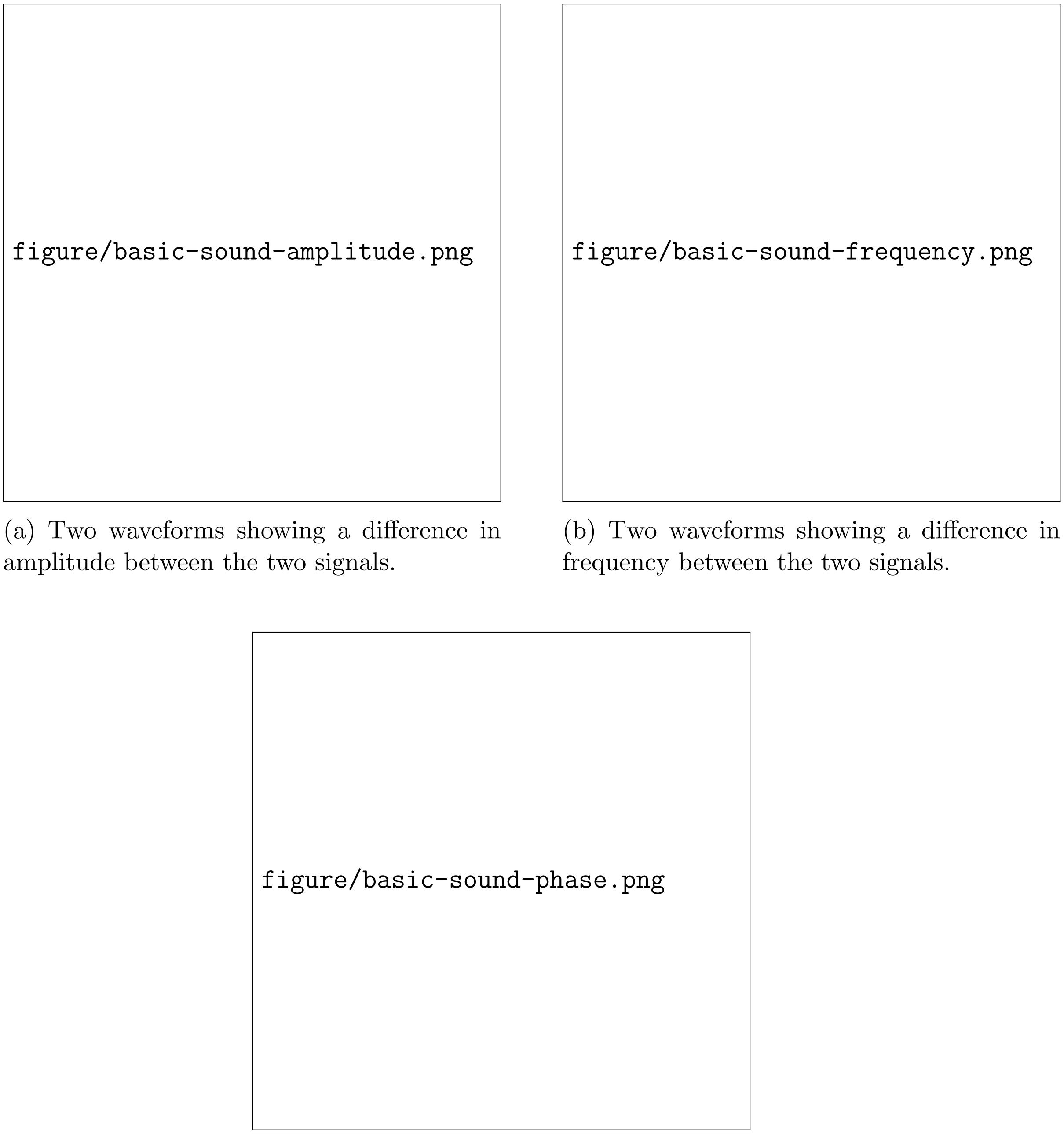
Sound itself, along with the ability to perceive it, is a remarkable phenomenon. Put simply, "sound" is the fluctuation of pressure in neighboring groups of particles over time. Most frequently sound is discussed in terms of the fluctuation of air pressure, because, as humans, we primarily receive sound into our ear canals through the medium of air, but sound can also travel through media of liquids and solids, or pass through any combination of the three.

There are primarily three components to sound: amplitude, frequency, and phase, which can be seen in Figure 1.1. In a simple sound wave, the amplitude of sound corresponds to the peak intensity of the high pressure (and the lowest pressure) portions of a sound wave (cf. Fig. 1.1a). The frequency corresponds to the rate at which the high and low pressure portions of the signal fluctuate between one another (cf. Fig. 1.1b). The phase of a wave is the location of the pressure level (relative to atmospheric pressure) in time (cf. Fig. 1.1c).

This latter characteristic - phase, while important when dealing with interacting waves from multiple sources, is often not taken into in speech science due to both its complexity and the fact that phase does not encode any speech information. The human auditory system primarily makes use of the other two characteristics of sound - amplitude and frequency.

1.1.1 Overview of Anatomy and Physiology of the Peripheral Auditory System

Prior to discussing the acoustic structure of speech, it is important to become familiar with the basic peripheral auditory system. The peripheral auditory system is generally grouped into three primary categories, the outer ear, the middle ear, and the inner ear (cf. Figure 1.2). The outer ear includes the pinna, the ear canal tube, and the tympanic membrane (ie. the eardrum). Air-transmitted sound vibrations, ie. pressure fluctuations, enter the ear canal through the opening at the pinna. These then travel along the canal to vibrate the tympanic membrane, which passes the energy to the middle ear. The middle ear includes the ossicles within the middle ear cavity. The ossicles are a chain of three very small bones leading from the tympanic membrane to the cochlea. The external sound vibrations hit the tympanic membrane, are passed along the ossicle chain, which then sends these vibrations to the inner ear.


figure/basic-sound-amplitude.png
figure/basic-sound-frequency.png
figure/basic-sound-phase.png

(a) Two waveforms showing a difference in amplitude between the two signals.

(b) Two waveforms showing a difference in frequency between the two signals.

(c) Two waveforms showing a difference in phase between the two signals.

Figure 1.1: Waveforms demonstrating difference in amplitude, frequency, and phase.

The inner ear is composed of the cochlea, the semicircular canals (and vestibule), and the auditory and vestibular nerves. The semicircular canals, vestibule, and figure/ear_anatomy.png

Figure 1.2: A diagram of the peripheral auditory system, including the outer ear, middle ear, and inner ear, up to the auditory nerve. (Image from **?**)

vestibular nerve don't play a part in audition (their primary function regards balance sensitivity). The cochlea, a hard 'shell' filled with fluid receives the vibrations passed along through the middle ear ossicles.[1] The vibrations are passed into the fluid cochlea and travel along the length of the cochlea, transmitting acoustic energy to cells that are able to detect the amplitude of the vibrations. Depending on the location of a cell along the length of the cochlea, it will pick up a different frequency of sound. The auditory nerve carries electrical impulses from these cells into the auditory cortex of the brain.

Of interest to this present study is the relatively simple outer ear. Typically, as described above, vibrations from the air will enter the ear canal through the opening at the pinna. Frequently, these vibrations take the form of human speech.

1.1.2 Acoustic Structure of Human Speech

The structure of human speech draws from a collection of individual sounds (phonemes) that are strung together, encoding words, sentences, and more abstractly, meaning. Each human language uses a subset of all possible phonemes. Each phoneme is either considered 'voiced' or 'unvoiced'. The acoustic properties of sounds in these two categories differ greatly.

Voiced speech is composed of narrow bands of acoustic energy, called harmonics, located along a frequency spectrum (cf. fig 1.3). In this sense, speech is considered a 'complex' sound, because it is composed of multiple, simultaneous frequencies.

[1] The peripheral auditory system uses gas, solid, and liquid media to transmit acoustic vibrations.

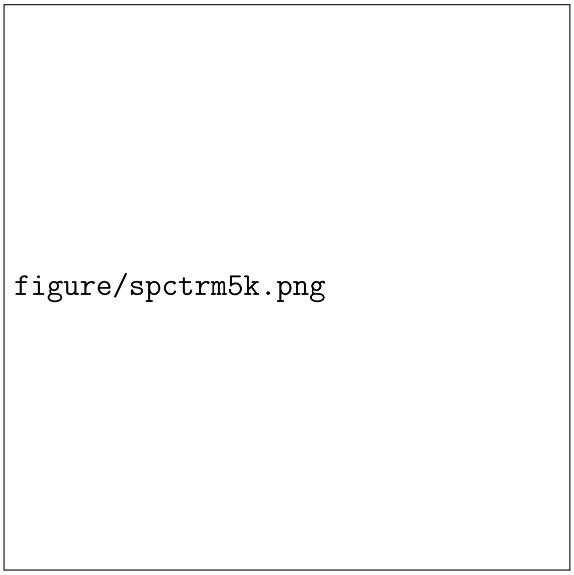


Figure 1.3: Spectrum of the middle of an /I/ vowel. Each ‘spike’ is a separate narrow band of frequency, called a harmonic.

Certain harmonics will be dampened by the vocal tract, leaving others relatively unfiltered. A group of neighboring harmonics containing more energy than other harmonics are called formants. The location, shape, and transition over time of these formants (among other more minor features) are what encodes speech information for voiced sounds. This can be easily visualized in a spectrogram (cf. Fig. 1.4).

For unvoiced speech, the information used to recognize and categorize the speech sound is likely found in either the turbulent frication generally centered in higher frequencies (cf. Fig. 1.5, although some of the information can be found in lower frequencies), or found in the voiced information in the transitions into and out of the sound (?).

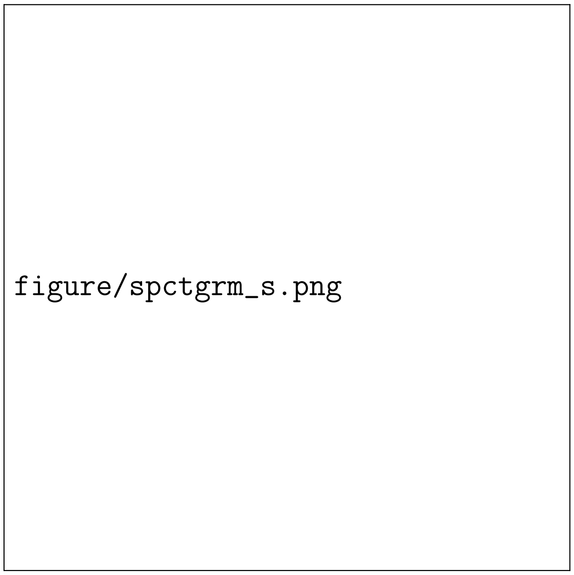


Figure 1.5: Spectrum of the initial /s/ in “citizen”. Zoomed to range of 0-8kHz for visualization of high frequency energy.

The human auditory system has the remarkable ability to (a) identify these sounds, which often only last from tens to a few hundreds of milliseconds in duration, (b) partition the stream of sounds into their respective words, and (c) string the words together into a sentence and pull meaning from it - all in real time. Nevertheless there are occasionally recognition errors, which can occur anywhere along this auditory chain. The ‘lowest’ level in this chain that errors occur is the recognition and identification of sounds. There are a host of reasons why this might occur, yet this report will focus on one of them - additive noise interfering with the speech signal.

1.1.3 Acoustics of Multiple Signals

As previously mentioned, we as humans use the amplitude and frequency of pressure fluctuations to perceive sound. When sound travels through a medium from source *A*, there is nothing that prevents these pressure fluctuations of source *A* from acoustically mixing with the pressure fluctuations originating from source *B*. For example, in Figure 1.6 source *A* produces a simple wave with a frequency of 100 Hz. Source *B* produces a simple wave with a frequency of 200 Hz. If the waves from the two sources reach each other and overlap, a single wave results that looks like that in Figure 1.6c. This sound wave now has two components - a tone at 100 Hz and a tone at 200 Hz.

As previously mentioned, phase does not play a significant role in human audition, but can affect a wave resulting from the addition of multiple sound sources. Say that source *B* produces instead a wave the same frequency and amplitude as the wave from source *A*, but they are completely 'out of phase', ie. the pressure value of the waves at any given time is in direct opposition (cf. Figure 1.8a). If these waves are combined, it results in a complete elimination of sound (cf. Fig. 1.7b).

In order for this complete negation to happen, the two waves need to be coming from opposite directions toward one another. Of course, it is not very often that two waves of the exact same frequency and amplitude, with exactly opposing phase, meet in such a way to completely negate. However, varying degrees of negation occur frequently due to phase. For example, if the 100 Hz wave from Figure 1.6 were combined with a 200Hz wave with a slightly shifted phase, a different wave would be produced, seen in Figure 1.8c.

The combination of waves from multiple sound sources increases greatly in complexity as the number of sources increases, and the sounds originating from the sources are complex (ie. containing multiple frequency elements), such as speech. Speech rarely occurs in isolation from from all external sound, yet we are still to largely understand speech in everyday environments; for example it is generally easy for humans to understand the speech of an interlocutor while sitting on a bench at a park.

The auditory system is actually quite skilled at identifying separate sources, even complex ones, like speech. Despite the shifted phase in Figure 1.8, the human auditory system would still be able to detect and identify two separate waves. While it undoubtedly plays a part, the differences in phase of combined signals does not normally completely negate a signal, nor render it unintelligible. It is for this reason, and the complexity of phase calculations, that most efforts to remove speech from noise ignore the phase component.

1.1.4 Difficulties of Speech in Noise

Nevertheless, there are still situations in which it is difficult to parse speech in noise. This is most often due to signals with energy at similar frequencies that overlap. The greater the amplitude of a signal at frequencies overlapping with those of speech, the more difficult the speech will be to understand. This can be visualized in the spectrograms of the sentence “A rich farm is rare in this sandy waste.” in Figure 1.9. In Figure 1.9a, the amplitude, or loudness, of the noise is well below that of speech; one would likely easily understand this speech. However, Figure 1.9b has a much greater noise level compared to speech.

This relationship between speech and any background noise is called the signal to noise ration (SNR). A complex signal with a higher signal to noise ratio (cf. Fig 1.9a) is generally easier to understand, because the amplitude of the speech (the ‘signal’ of interest) is much greater than that of the noise. Consequently a lower SNR (cf. Fig. 1.9b) results in speech that is more difficult to understand, because the amplitude of the speech is close to - or below - the amplitude of the background noise.

This poses a problem for human listeners, but generally is more difficult to deal with in automatic speech recognition (ASR), since the electronic device does not contain a highly-skilled, built-in auditory cortex. There are a number of ways which have been proposed to deal with noise in a speech signal, both for human and automatic speech recognition. These will be discussed further in Chapters 3**??** and 4**??**.

1.2 Overview of Dissertation

This report aims to explore a novel method of human speech perception and automatic speech recognition (ASR) in noisy environments. The method proposes that speech be recorded from the inside of the ear canal of the speaker, and slightly transformed, sent to the human hearer or the computer receiver for recognition. By collecting speech from the ear, it allows for usage of the human skull and adjacent tissues to passively filter out the noisy environment, leaving only - or mostly - the human speech carrying the intended message.

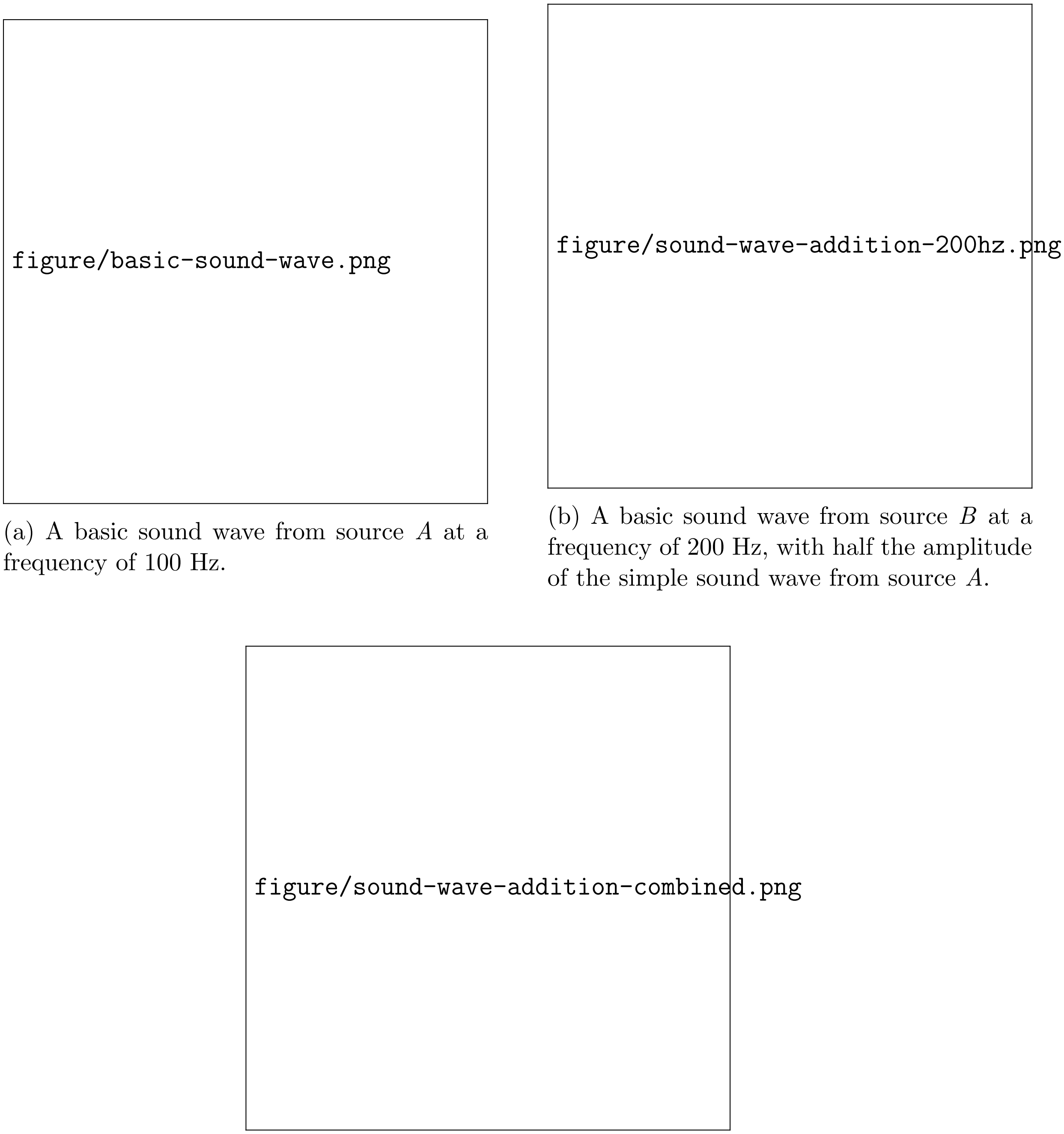
The intention of this study is to determine a) if recording human speech from the inside of the ear canal can significantly reduce background noise in a signal, b) if intelligible speech, suitable for communication, can be collected from the inside of the ear canal, c) if humans find speech recorded from the ear canal more intelligible than speech in noisy conditions, and finally d) if ASR systems are able to recognize speech recorded in the ear canal with greater accuracy than speech recorded in noise.

Since currently there is no established corpus of data that contain speech recorded from the inside of the ear canal together with speech simultaneously recorded from the mouth in noisy environments, it was necessary to record speakers in this environment and create a new corpus. The theory behind the acoustics of recording speech from the ear canal, as well as the process for developing this corpus are described in Chapter 2**??**, along with a discussion of the recorded speech. Chapter 3**??** will outline a human perception experiment, tasking listeners with the transcription of various sentences of speech recorded at the mouth in noise in noise, and recorded from the ear. Chapter 4?? describes the use of this same speech with an ASR system, and its recognition performance. Chapter 5?? will summarize the previous chapters and engage in an overall discussion of the implications of the results, the limitations of the present experiments and methods, and suggestions for future research direction.

figure/spctgrm1k.png

(a) Zoomed to the 0-1kHz range for better visualization of harmonics.

figure/spctgrm5k.png


figure/basic-sound-wave.png
figure/sound-wave-addition-200hz.png
figure/sound-wave-addition-combined.png

(a) A basic sound wave from source *A* at a frequency of 100 Hz.

(b) A basic sound wave from source *B* at a frequency of 200 Hz, with half the amplitude of the simple sound wave from source *A*.

(c) The resulting complex wave from the combination of the wave from source *A* and source *B*.

Figure 1.6: Demonstration of the combination of two waves of different frequency and amplitude.

figure/wave-out-of-phase.png

(a) Two sound waves with frequency of 100 Hz and the same amplitude, completely out of phase.

figure/wave-out-of-phase-added.png

(b) The sound wave resulting from the combination of the two out of phase waves in Fig. 1.8a.

Figure 1.7: Demonstration of the combination of two completely out-of-phase waves with the same amplitude and frequency.

figure/basic-sound-wave.png

(a) Two sound waves with frequency of 100 Hz and the same amplitude, completely out of phase.

figure/sound-wave-addition-200hz-shifted.pn

(b) The sound wave resulting from the combination of the two out of phase waves in Fig. 1.8a.

figure/sound-combined-shifted-phase.png

(c) The sound wave resulting from the combination of the two out of phase waves in Fig. 1.8a.

Figure 1.8: Demonstration of the combination of the same two waves as in Fig. **??**, where the phase of the 200 Hz wave in Fig. 1.6b was shifted slightly (cf. Fig. 1.8b)

figure/signal-SNR-intro-high.png

(a) A sentence spoken with a low level of background noise, resulting in a *high* SNR.

figure/signal-SNR-intro-low.png

(b) A sentence spoken with a high level of background noise, resulting in a *low* SNR.

Figure 1.9: Waveforms and spectrograms of the sentence “A rich farm is rare in this sandy waste”.

REFERENCES