COMPUTATIONAL DIFFERENCES BETWEEN WHISPERED AND NON-WHISPERED SPEECH

BY

BOON PANG LIM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor Mark A. Hasegawa-Johnson, Chair
Adjunct Professor Richard W. Sproat, Co-chair
Assistant Professor Ryan K. Shosted
Professor Stephen E. Levinson
Associate Professor Steven S. Lumetta

ABSTRACT

Whispering is a common type of speech which is not often studied in speech technology. Perceptual and physiological studies show us that whispered speech is subtly different from phonated speech, and is surprisingly able to carry a tremendous amount of information. In this dissertation we consider the question: What makes whispering a good form of communication? We examine the differences between normal phonated speech and whispered speech, and gauge the effectiveness of state-of-the-art speech recognition algorithms at recognizing whisper. Our perceptual experiments add to the literature on the intelligibility of whispered speech. Comparisons with ASR results yield interesting insights into the differences between the two systems.

A method for building speech recognizers for whispered speech using limited whispered speech data is proposed and evaluated. Our approach effectively performs speaker-adaptation for whispered speech acoustic models without needing whispered speech from the target speaker. Results show improvement over the standard speaker-independent models. Our work opens up additional avenues for research, which are outlined in the conclusions.

To friends, family and teachers

ACKNOWLEDGMENTS

It is often said that no significant work can be accomplished alone — in this original work, while I do not have a magnum opus, I am indebted to many people for having contributed to it, either directly or indirectly. I owe thanks foremost to my advisers, Professors Richard Sproat and Mark Hasegawa-Johnson. Richard has been most supportive and tolerant of my rather odd work habits and has contributed many ideas in discussions that have helped to make this work more complete. In fact, this work came about as a suggestion of his after one of our weekly discussions — he helped me find a useful topic while I was languishing in graduate school. After Richard's move to Oregon, I had the blessing of working with Mark, who supervised me closely and oversaw the development of my new adaptation algorithm. His immense expertise on speech recognition systems has given me new insight into how these systems work or should work. Many fruitful ideas have come from our discussions.

My thesis committee, Professors Stephen Levinson, Ryan Shosted and Steve Lumetta, have been extremely supportive throughout my graduate career. I took two very insightful classes on speech processing from Professor Levinson. These classes have helped to lay my foundations for the subject matter. Professor Shosted has been very helpful and willing to discuss my work and its ramifications, often providing good references for me to follow up. Professor Lumetta, with his computer engineering background, very willingly offered to be on my committee when a technicality about its member composition cropped up. Thanks to Professor Jennifer Cole, who taught me acoustic phonetics. To these wonderful teachers I remain in deep gratitude.

Life in graduate school would not be tolerable if not for my fellow cubicle-sharers in the Statistical Speech Technology Group. To Sarah King, Lae-Hoon Kim, Yoon Su-Yeon, Harsh Vardan Sharma and Jui-Ting Huang, I owe thanks. They, especially Sarah and Lae-Hoon, have provided me many

opportunities for interesting discussions. Their constant presence along with their expertise in their own subject matter has proved invaluable, as they were immediately accessible encyclopedias on speech technology as well as a source of warm friendship. I owe special thanks to Professor Stephen Bishop, who gave me expert counsel on my options in graduate school and also extended his friendship despite his high-ranking position as associate head. I also owe thanks to the administrative staff in the department who have helped me at various times, as well as to the Publications Office staff for their thorough and detailed checks. Also I owe thanks to my friends from Singapore, Jamie Teo, Teck-Leong Tan and Zhun-Yong Ong, who supported me in my most difficult times of despair.

I credit Dr.'s Li Haizhou and George White for sparking my interest in speech recognition. Both of them started me on the path to learning about speech while I was working at the Institute for Infocomm Research, and they offered me many opportunities for working on interesting projects.

Last, but definitely not least, I thank my family for their infinite understanding and support. I thank my dear sister, who has helped me in my many times of need, offering counsel, support and prayer. I also thank my mother, who supported me every single day, even performing menial but important things like packing my lunch in the last months of writing this dissertation.

I owe much to all these people, and more to many others. I thus dedicate this work and whatever meager contribution from it to these people: to teachers, friends and family, without whom I would have accomplished nothing at all.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

LIST OF ABBREVIATIONS xi

CHAPTER 1 INTRODUCTION AND BACKGROUND 1
1.1 Introduction: The Trouble with Whispering 1
1.2 Speech Communication 2
1.3 Organization of the Dissertation 24

CHAPTER 2 WHISPERED SPEECH 26
2.1 Production of Whispered Speech 26
2.2 Acoustics of Whispered Speech 29
2.3 Perception of Whispered Speech 31
2.4 Applications of Speech Technology to Whispered Speech . . . 38

CHAPTER 3 AUTOMATIC SPEECH RECOGNITION 45
3.1 Overview of LVCSR 45
3.2 Acoustic Pattern Recognition Using Hidden Markov Models . 53
3.3 Speaker-Independent Acoustic Modeling 57
3.4 Speaker-Adaptation Techniques 58
3.5 Summary 62

CHAPTER 4 WHISPERED SPEECH CORPORA 63
4.1 The Whispered TIMIT Corpus 63
4.2 The Whispered Modified Rhyme Test Corpus 66
4.3 Acoustic Analysis 66

CHAPTER 5 THE PERCEPTION OF WHISPERED SPEECH . . . 75
5.1 Experiment Design of the Whispered Modified Rhyme Test . . 75
5.2 Human Perception in the Whispered Modified Rhyme Test . . 77
5.3 Machine Recognition of wMRT Sentences 86
5.4 The Effect of Context in Communication 89
5.5 Discussion of Results 94

CHAPTER 6 RECOGNITION OF WHISPERED SPEECH 97
6.1 Accuracy of Speaker-Independent Acoustic Models 97
6.2 Implementing Eigenvoices in HTK 99
6.3 Speaker Adaptation with Normal and Whispered Speech . . . 102
6.4 Adapting Speaking Style and Accent Using CMLLR 104
6.5 Acoustic Model Adaptation with Limited Whisper Data 105
6.6 Summary 111

CHAPTER 7 CONCLUSION 113
7.1 Summary of Completed Work 113
7.2 Future Work 114
7.3 Conclusion 117

APPENDIX A FORMANT MEASUREMENTS 118

REFERENCES 121

LIST OF TABLES

1.1 Breaking down the process of speech communication. 4
1.2 Consonants in English. 14

4.1 Speakers in first phase collection (collected at NUS). 64
4.2 Speakers in second phase collection (collected at UIUC). . . . 65
4.3 List of word-initial question sets in the Modified Rhyme Test. 67
4.4 List of word-final question sets in the Modified Rhyme Test. . 68

5.1 Per listener wMRT identification accuracies. 77
5.2 Per speaker wMRT identification accuracies. 78
5.3 Human performance for word-initial and word-final consonant recognition. 79
5.4 Most common confusions in human wMRT perception. 80
5.5 Perceptual error confusions for stops. 83
5.6 Perceptual error confusions for nasals. 83
5.7 Perceptual error confusions for fricatives and affricates. . . . 84
5.8 Accuracy of voicing feature transmission computed from wMRT result. 85
5.9 Overall identification accuracy for different genders. 86
5.10 Machine and human performance for whispered and unwhispered speech recognition. 88
5.11 Error confusions for stops in ASR wMRT. 90
5.12 Error confusions for nasals in ASR wMRT. 90
5.13 Confusions for fricatives and affricates in ASR wMRT. 91
5.14 Accuracy of voicing feature transmission computed from ASR wMRT result. 91
5.15 Estimate of information carried by voicing in speech. 95

6.1 Word recognition accuracy across different models and datasets. 99
6.2 Delta-word recognition accuracy – improvement over SI baseline for different speaker adaptation methods (normal acoustic models). 103

6.3 Delta word recognition accuracy – improvement over SI baseline for different speaker adaptation methods (whispered acoustic models). 104
6.4 Relative WER reduction – comparison of speaker adaptation methods (in %). 104
6.5 Word recognition accuracies obtained from adapting accent and speaking style with CMLLR. 105
6.6 Word recognition accuracies of speaker-dependent whisper acoustic model produced from normal speech of said talker. . 111

A.1 Formant frequencies for /i,a,u/ in fluent speech (wTIMIT-US). 119
A.2 Formant frequencies for /i,a,u/ in fluent speech (wTIMIT-SG). 120

LIST OF FIGURES

1.1 Components in speech understanding 2
1.2 Sounds of the world's languages. 7
1.3 The human speech apparatus. 10
1.4 The larynx and vocal folds. 10
1.5 The source filter model of speech production 17
1.6 Cutaway section of the ear 18
1.7 Model of the Basilar Membrane 18
1.8 Intermediate processing leading up to the auditory cortex. . . 20
1.9 The acoustic waveform and narrowband spectrogram for both normal and whispered speech 23

2.1 Basic framework for reconstructing voice. 43

3.1 Speech recognition by humans versus machine. 47
3.2 Example of an embedded word graph 50
3.3 A signal processing front end for speech recognition. 51
3.4 Diagram of a three-state left-to-right HMM. 54

4.1 Spectrograms of normal and whispered speech. 69
4.2 Log-spectral plots for whispered and normal speech. 71
4.3 Change in formant frequencies going from normal to whispered speech. 73
4.4 Average increase in phone length due to whispering. 74

5.1 Screenshots from testing program. 76
5.2 Categories of errors found in perceptual wMRT. 81
5.3 Example of a regular grammar for a wMRT question set. . . 87
5.4 Categories of errors found in ASR wMRT. 88
5.5 Filtering approach to estimate information transmitted. 92
5.6 Reducing entropy of the language with increased context. . . 94

6.1 Illustration of eigenspace mapping approach 107
6.2 Illustration of projection-based eigenvoice mapping procedure 108

7.1 Continuous variable duration hidden Markov models. 116

LIST OF ABBREVIATIONS

ASCII	American Standard Code for Information Interchange
ASR	Automatic Speech Recognition
BNF	Backus-Naur Form
CFG	Context-Free Grammar
CMLLR	Constrained Maximum Likelihood Linear Regression
CMU	Carnegie-Mellon University
CTM	Close-Talking Microphone
CV	Consonant-Vowel
CVC	Consonant-Vowel-Consonant
DAM	Diagnostic Acceptability Measure
DBN	Dynamic Bayesian Network
DWP	Discrete Wavelet Packet
DWT	Discrete Wavelet Transform
DFT	Discrete Fourier Transform
DMOS	Diagnostic Mean Opinion Score
DRT	Diagnostic Rhyme Test
FIR	Finite Impulse Response
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HMM	Hidden Markov Model

HTK	HMM Toolkit
IEEE	Institute of Electrical and Electronics Engineers
IPA	International Phonetic Alphabet
JMLS	Jump Markov Linear System
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
LTAS	Long-Term Average Spectrum
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum A Posteriori
MELP	Mixed-Excited Linear Prediction
MEMM	Maximum-Entropy Markov Model
MFCC	Mel-Frequency Cepstral Coefficients
MLED	Maximum-Likelihood Eigen-Decomposition
MLLR	Maximum-Likelihood Linear Regression
MRT	Modified Rhyme Test
NIST	National Institute of Standards and Technology
NUS	National University of Singapore
PCA	Principal Components Analysis
PDF	Probability Density Function
PHS	Personal Handphone System
PPCA	Probabilistic Principal Components Analysis
RASTA	Relative Spectral Analysis
SD	Speaker Dependent
SI	Speaker Independent
SNR	Signal-to-Noise Ratio
SRI	Stanford Research International
STFT	Short-Time Fourier Transform

TIMIT	Texas-Instruments, Massachusetts Institute of Technology
TRAPs	Temporal Patterns (feature extraction method)
UIUC	University of Illinois at Urbana-Champaign
WER	Word Error Rate
wMRT	Whispered Modified Rhyme Test
wTIMIT	Whispered TIMIT

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Introduction: The Trouble with Whispering

Human speech is a natural mode of communication that is well studied but perhaps not well understood. At least, engineers have so far failed to apply our understanding to practice — automatic speech recognition (ASR) techniques fall far short of human performance. Simple changes in the recording environment and in speaking style will deteriorate the performance of existing state-of-the-art techniques. In this dissertation we are concerned with one simple but devastating (to performance), yet not often studied, deviation to normal speech — whispering.

Several pressing problems persist with the study of whispered speech recognition, the foremost of which is the lack of a large, systematic, publicly available corpus for study. This in turn presents an interesting problem: Can algorithms be designed that will work with whispering, using a reasonable amount of normal speech and minimal amount of whispered speech for training? How is whispered speech different from normal speech, and how do these differences manifest in the acoustics and affect the performance of speech recognition? More fundamentally, how good is whispering itself as a channel for conveying spoken information, even between human speakers and listeners?

All of these problems, while not necessarily resolved, are addressed by the data and experiments in this dissertation. Before we do so, it is important to understand in greater detail the nature of speech and whisper as we understand them today.

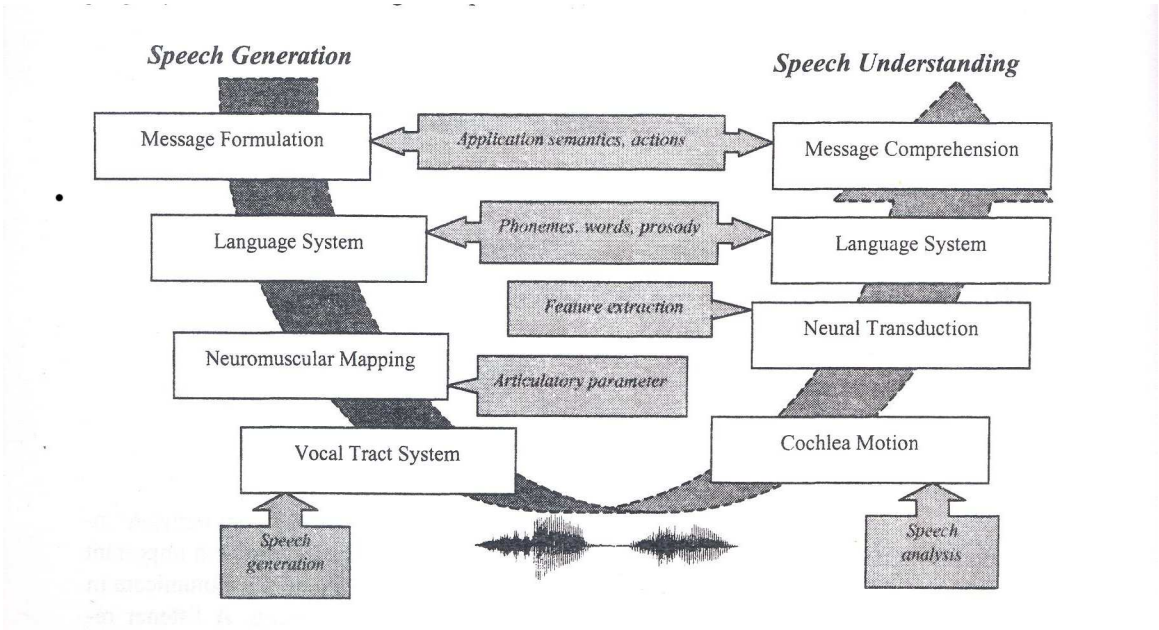


Figure 1.1: Components in speech understanding (taken from [4]).

1.2 Speech Communication

Speech is a primary mode of human communication that emerged late in human evolution - the facilities for the production and perception of communicative speech can be thought of as having evolved from pre-existing anatomy evolved for breathing [1] and eating. Although it is unclear if the eating apparatus further coevolved as anatomical changes to permit speech developed,[1] most definitely there is a multiplexing of several, non-speech related functions on the same physiological setup. The impact of this evolution is tremendous, especially in terms of the human articulatory apparatus; at a first glance it appears to be nothing at all like an engineering system designed for the sole purpose of communication.

The complex process of communicating messages with spoken language may be broken down into stages. As shown in Figure 1.1, communication begins with the formulation of the message, and goes through several stages, propagating through the acoustic medium, and goes up the message analysis

[1]Comparative study of primate skulls [2] provides evidence that properties allowing speech articulation, such as the lowering of the larynx and compaction of muscle structures to the base of the skull thus increasing the mobility of the tongue, co-occurred and perhaps were a result of man taking an upright posture. At the same time, lowering of the larynx increases the danger of food going into the breathing airways and could be seen as counterproductive to more efficient eating [3].

chain, ending with message comprehension. Prior to actual articulation of the message, we may have cognitive processes that decide what words to communicate in the first place, and after receiving the sounds and decoding them in the ear there are linguistic and cognitive processes that perform actual understanding of the message, where semantics and pragmatics come into play. However, if we are more concerned with what goes on close to the acoustic medium itself, in engineering terms we are left with basically the articulatory apparatus (transmitter), a medium over which the sound propagates (channel), and apparatus that hears and performs initial decoding of the incoming speech sounds (receiver). The corresponding apparatus for transmitting and receiving speech, the fields of study associated with them and the machine technologies that have been developed can be broken down in this manner as illustrated by Table 1.1.

The rest of this section will provide the reader with an understanding of key ideas in the related sub-fields in linguistics (phonetics) [5, 6], psychophysics [7] and engineering (speech technology) [8], starting with the source of the speech signal, then examining in detail the recognition and identification of speech, and finally looking at how speech sounds manifest in the acoustic waveform.

Just as it is suboptimal to design the transmitter of a communicative system independently of the channel and the receiver, and preferable to consider the limitations on the entire link at once [9], it would seem myopic to study speech without an understanding of all three domains - articulatory (speech production), perceptual (cognition), and acoustic (waveform and spectrogram) [6]. The linguistic units (e.g., phones, syllables and words) that make up a spoken language manifest in all three domains:

- In speech production, they can be described in terms of the movement and neural control of the speech articulators. The *articulatory correlates* are articulatory kinematic variables correlated with the absence or presence of the linguistic unit.

- In speech perception, they can be described in terms of *perceptual correlates* — the stimulus-response in the auditory nerve and tonotopic excitations within the auditory cortex.

- In acoustic phonetics, they can be described in terms of *acoustic cor-*

Table 1.1: Breaking down the process of speech communication.

Engineering Concept	Physiology Physical Medium	Field of Linguistic Study	Engineering Model
Transmitter	Human speech apparatus	Articulatory Phonetics	Speech Synthesis
Channel	acoustic waveform	Acoustic Phonetics	Speech Coding
Receiver	the ear, auditory nerve, auditory cortex	Speech Perception	Speech Recognition

relates — these are any acoustic measurements that correlate with the absence or presence of the linguistic unit, examples of which are the location and transition of the formants and spectral shape of noise-like regions.

Phonemes: Consonants and Vowels

The study of speech sounds is known as phonetics [10]. Despite the wide variance and sheer number of languages available in the world today, the basic sounds of any language can be categorized into perceptually similar short segments of speech. There are many phonological theories which concern this, but the more popular ones involve the *phoneme*. An unwavering definition of *phoneme* itself is a point of contention even among linguists [11], but we shall try to provide a useable definition here that we as speech technology engineers can use.

- The phoneme is a readily identifiable unit of speech — usually it is a minimally distinctive unit of sound in a language. It is also related to the distribution of letter-sounds in alphabetic languages. To the layman, the phoneme appears to correspond to basic speech sounds that make up a word (for instance /k ae sh/ in ARPAbet [12] or /k æ ʃ/ in IPA [13]). Although it need not be a natural construct of all languages, nor is it a necessary part of all phonological systems and theories, it appears that a phonemic inventory exists for all languages [11].

- In English, phonemes can be generally categorized as either vowels or consonants, each of which have different articulatory, acoustic and perceptual properties [14]. Consequently, they have acoustic and perceptual correlates — the identity of a phoneme can be signaled by features in the acoustic waveform (e.g. CV transitions), or by excitation of specific regions in the auditory cortex (i.e. a tonotopic mapping). Some phonologists [15] even go so far as to suggest a fourth correlate: a mental representation of phoneme as it is to be produced, although there is no compelling reason why this cannot be the same as the perceptual representation of a phoneme.

- Vowels are produced with a relatively open vocal tract, and thus have a resonant sound [5]. In contrast, consonants are produced with a narrow constriction, sometimes causing turbulence in the airflow (fricatives); in some cases there is a complete stoppage of the airflow (stops).

- Phonemes string into words, usually in a consonant-vowel (CV) or consonant-vowel-consonant (CVC) pattern. Each language's phonemes may manifest differently (i.e. *allophones*) depending on their location in the word. (E.g. in English, /t/ is accompanied by a puff of air — i.e. *aspirated* — when it occurs at the start of words, but it is not aspirated when it is at the end) [14].

- A specific language may have its own unique set of phonemes and allophones. Similar sounding words can often be distinguished by the difference in just one phoneme; such pairs of words (e.g. bash /b æʃ/ vs. dash /d æʃ/, differ only in *place of articulation* of the first phoneme) are called *minimal pairs*. Furthermore, the possible combinations of sounds in a language tend to be severely limited; for instance, in Mandarin, words tend to be CV or CVC in nature, with the final consonant, excepting "retroflex finals," either an /n/ or /ŋ/ [16]. Grammatically allowed concatenations of phonemes within a language are governed by its *phonotactics* [17].

- In actual speech, the movement of the articulators and the control of the glottis are not strictly synchronized - while perceptually we can easily identify whether a consonant follows a vowel or vice versa, within the acoustic signal it is often hard to find a strict time boundary separating a pair of contiguously articulated phonemes. Rapid movement of the articulators in natural speech results in co-articulation - which can be thought of as a kind of inter-symbol interference, manifesting as different acoustic observations for different contexts preceding or following a phoneme. Consequently, it is useful to consider the characteristics of syllables rather than phones, as all languages are syllabic, and the identity of CV sequences is largely manifest in the first and second formant transition [18].

The study of phoneme manifestation in the acoustic waveform or spectrogram (i.e. the acoustic correlates) falls under the sub-field of acoustic

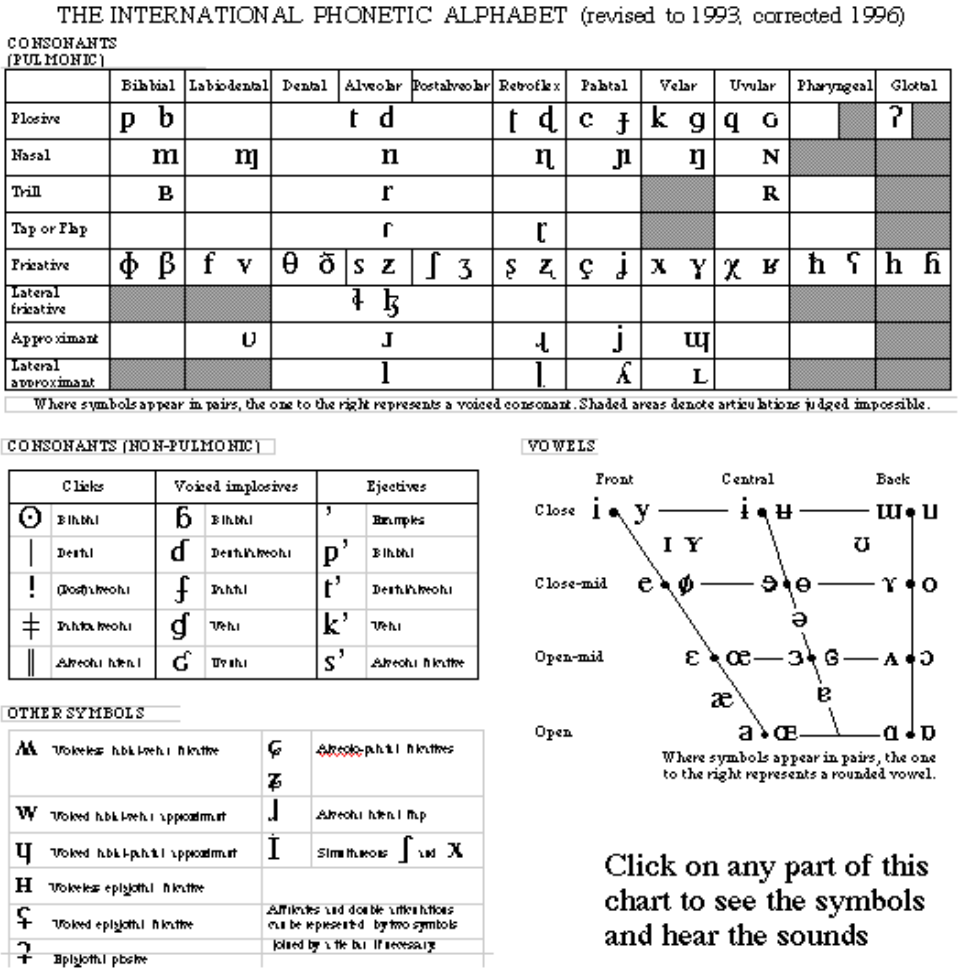


Figure 1.2: Sounds of the world's languages (taken from Peter Ladefoged's website [20]).

phonetics [19]. The sounds of the world's languages can be organized by their properties, as shown in the International Phonetic Association (IPA) chart in Figure 1.2. The chart shows how consonants and vowels manifest in the articulatory domain: vowels are organized by the height and position of the tongue; consonants are organized by how they are articulated and where they form the smallest constriction. This chart is readily available from Peter Ladefoged's website [20] or from the website of the IPA (www.langsci.ucl.ac.uk/ipa).

Distinctive Feature Theory

The phonemes in English may be distinguished by their distinctive features [21]; these are binary properties or features which may be either present or absent during their production. These features could be arranged hierarchically, but more often they are thought of in terms of two main groups:

- The articulator-bound features describe presence or absence of position and or movement of the speech articulators. They can be further organized into three major groups, depending on whether they refer to movement of articulators in the oral cavity, movement in the pharyngeal cavity or muscle stiffness in the larynx. These features include:
  - movement in oral region — round, anterior, distributed, lateral, high, low and back
  - movement in pharyngeal region — nasal, advanced tongue root, constricted tongue root, spread glottis and constricted glottis
  - surface stiffness — stiff vocal folds, or slack vocal folds
- The articulator-free features describe features or properties without referring to the speech articulators, and these tend to be perceptual in nature. These include features such as `[consonantal]`, `[vocalic]`, `[sonorant]`, `[strident]`, and `[continuant]`.

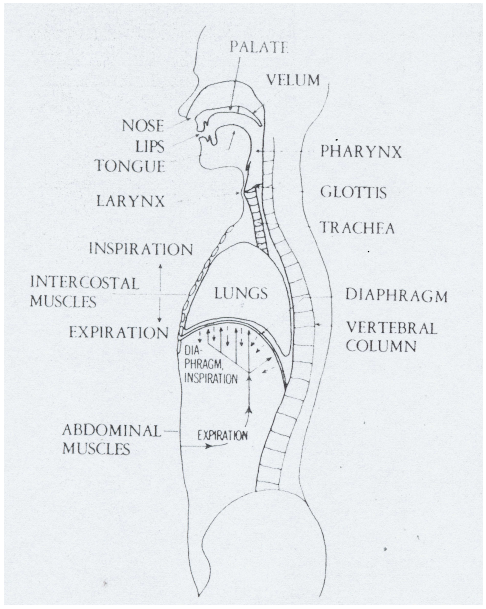
A thorough definition of all of these features is beyond the scope of this writing, and the reader may refer to the excellent writings of Chomsky and Halle [21], or modern textbooks on phonetics [10]. However, it must be noted that all of the features, like phonemes themselves, can be thought of in terms of their acoustic and perceptual correlates. The *landmark* can be thought of as a sub-feature of a phoneme, in the sense that it is a linguistic unit that is part of a phoneme, much as a phoneme can be part of a syllable or a word, except that the "part of" relationship is largely temporal in nature. As these are linguistic units, they can manifest in all three domains (articulatory, perceptual and acoustic) in different ways. In particular, the acoustic landmarks used in speech recognition [22] are thought of by researchers as perceptual correlates of distinctive features. We next look at how these phonemes are produced, perceived and physically manifest in acoustics.

1.2.1 Speech Production

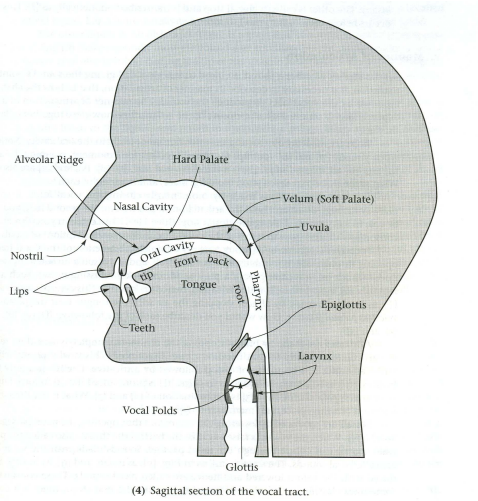
Speech production involves many parts of the human upper body. Stevens [5] divides these into three parts: the system below the larynx, the larynx and the surrounding structures, and the structures and airways above the larynx. This is illustrated with Figure 1.3(a). The system below the *larynx* consists of the respiratory structures: the abdomen, lungs and trachea. The lungs have a fractal-like quality to them — starting from the trachea these branch out into smaller bronchi, eventually branching into sac-like alveoli where gaseous exchange takes place during breathing. It is here that the process of producing speech begins, as the respiratory system itself serves as an energy source for phonated speech [6]. Similar to the expiration phase of breathing, the diaphragm moves upwards and compresses air out of the lungs, but the thoracic muscles and diaphragm contract in a controlled manner to maintain a constant rate of decrease in lung volume and nearly constant subglottal pressure [23]. This creates an airflow which passes through the natural constriction formed by the laryngeal structures.

Figure 1.3(b) shows a close-up diagram of the larynx and vocal tract. Air-flow passes through *vocal cords* at the *larynx*, into the cavities formed by the wall of the mouth (*oral cavity*) and breathing passage via the nose (*nasal cavity*) [6]. The *levator veli palatini* muscle attaches to the soft palate — its contraction raises the soft palate and seals off the nasal passage from the oral cavity; lowering the soft palate allows air-flow through the nasal passage giving rise to a nasal sound. The *vocal tract*, consisting of the oral and nasal cavities, acts somewhat like a resonant acoustic waveguide with a closed boundary near the *vocal cords*, and an open boundary at the lips. The resonant sound radiates outward from the lips as a pressure field [8], where it is picked up and perceived by human listeners as speech.

Normally spoken speech is phonated — this is due to action by the vocal folds. They situate in the middle of the larynx, as shown in Figure 1.4(a), and control air-flow into the vocal tract. A top-down view of the larynx showing the vocal folds and glottal configuration is shown in Figure 1.4(b). This is the view that is obtained through stroboscopy or laryngeoscopy. The abduction (pulling apart) and adduction (pushing together) of the vocal folds, and consequently opening and closing of the glottis, is controlled by the vocalis, crioco-arytenoid and inter-arytenoid muscles. The crico-arytenoid

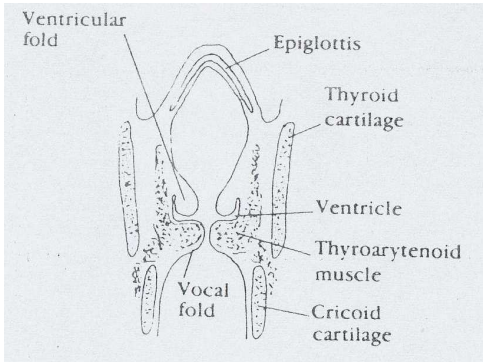


(a) Overview of lungs, larynx and vocal tract (from Stevens [5])

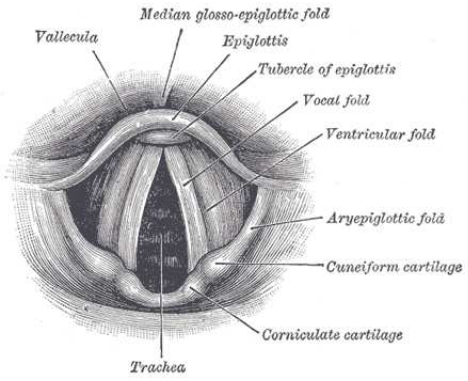


(b) Close-up view of vocal tract (taken from [24])

Figure 1.3: The human speech apparatus.



(a) Lateral view of the larynx (from Stevens [5])



(b) Illustration of the vocal folds, as possibly viewed from the top using a laryngeoscope (from Wikipedia Commons)

Figure 1.4: The larynx and vocal folds.

muscles can be held tense or lax, respectively opening or closing the air passage between the lungs and trachea to the mouth. The shape of the glottal opening or so called "glottal configuration" can take several forms, depending on which part of the folds are adducted.

In phonation, airflow through the vocal folds causes them to vibrate, opening and closing the glottis rhythmically. This aerodynamically driven motion [25] rhythmically closes and opens the glottis, effectively releasing air into the airways above the larynx in short pulses. The vibration typically happens at around 100 Hz for the adult male speaker and 200 Hz for the adult female speaker. This vibration manifests acoustically as F0, which is one of the major components of pitch. Slowly varying rates of vocal fold vibration give rise to pitch contours, which in turn give rise to intonation patterns in normally voiced sentences. Phonation at the glottis also breaks up the entire acoustic system at the glottis, effectively decoupling the subglottal and supraglottal mechanisms — in most acoustic treatments the glottis and other structures below including the lung and trachea are treated as a pitched acoustic source that excites the oral cavity.

The oral cavity itself acts like an acoustic waveguide that varies in cross-sectional area along its length [26, 27]. Its shape is largely determined by the movement of the jaw, tongue and lips. The position of the tongue tip separates the cavity into front and back portions — leading to a simplification with many treatments of the oral cavity which model it as a two-tube acoustic system. Such a system has natural resonances, *formants*, that are present in the resulting sound. The lowering of the velum can allow an additional route for airflow to escape out of the oral cavity via the nasal passage; this generally gives rise to the voice quality we know as *nasality*.

1.2.2 The Acoustic Theory of Speech Production

Engineers [28, 8, 5] have long modeled the production of speech using a *source-filter model*. The acoustic theory considers the body parts involved in speech production "before" the glottis (inclusive of the vibrating folds, trachea and lungs) as a source of excitation that drives a time-varying acoustic tube that acts as a filter. The shape of this tube is determined by the placement and position of the speech articulators.

Vowel Production and Two Tube Model

Vowels are resonant sounds produced with a relatively open configuration of the vocal tract with a continuous airflow. Under such conditions, airflow is non-turbulent and laminar: that is, air-flow is largely "parallel" along the vocal tract [29]. In the absence of nasalization, the acoustics of the vocal tract are very accurately modeled by treating it as a single tube of varying girth along its length. The area function $A(x,t)$ is defined as the cross-sectional area at time t along the vocal tract, at the position x centimeters away from the glottis. The acoustics of this system can be modeled by the Webster equation [30],

$$\frac{\partial^2 p}{\partial x^2} + \frac{1}{A(x,t)}\frac{\partial p}{\partial x}\frac{\partial A}{\partial x} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}, \tag{1.1}$$

where $p(x,t)$ is the pressure along the vocal tract. The resonant frequencies of the tube can be solved by discretizing the above partial differential equation and applying a gradient descent search technique. These resonances can be clearly observed within the speech spectrogram — they are called the formants — and numbered in ascending order from the lowest frequency formant to the highest one.

This model can be further simplified by simply considering just the point of the narrowest constriction; this is typically determined by the shape of the tongue which sets the location of the narrowest constriction, forming this with the tongue blade in the front vowels, and with the body for the back [19]. This constriction divides the oral cavity into a front (nearer the lips) and a back cavity, and the acoustics may be crudely modeled as two conjoined cylindrical tubes, each tube approximating acoustics of each cavity [5]. The analytic solution to such an acoustic model gives natural resonances that mimic very closely what is seen in measurements of actual vowels. The height of the tongue determines the so called height of the vowel produced; this shows up in the speech signal as the frequency of the first formant `F1` - the lowest natural resonance of the vocal tract. The location of the constriction determines the *frontness* or the *backness* of the vowel; resonances in the front cavity are associated with the second formant `F2`.

Production of Consonants

Consonants are speech sounds articulated with complete or partial closure of the vocal tract. The most common taxonomy of consonants in English is illustrated by Table 1.2 [24]. They are arranged from left to right by their *place of articulation* - the point of the narrowest constriction or greatest turbulence in the airflow within the vocal tract during their production. Going from the front to the back of the mouth, these places of articulation are

- labial — at the lips
- labiodental — the constriction involving the upper teeth and the lower lip
- interdental — with tip of the tongue placed between the teeth
- alveolar — with tongue tip near the alveolar ridge
- palatal — with tongue blade near the hard palate
- retroflex — with tongue tip near the hard palate
- velar — with tongue body near the velum (soft palate)
- pharyngeal — at the pharynx
- glottal — near the glottis

The consonants are categorized vertically in the consonant chart, according to their *manner of articulation* - the way in which they are produced. These categories include

- stops/plosives - in which there is a temporary but complete obstruction in the vocal tract, during which airflow continues to build up pressure at the back of the constriction. The plosives are characterized by four stages of production [14]:
  - closing phase - during which the movement of the articulators bring about the complete stop of the airflow through the vocal tract.

Table 1.2: Consonants in English. This is adapted from [24]. The symbols we use here are IPA, and the entries in parentheses are ARPAbet representations that are used in our machine pronunciation dictionaries.

Manner \ Place	bilabial	labiodental	interdental	alveolar	palatal	velar	glottal
Stop/Plosive	/p/ (P) /b/ (B)			/t/ (T) /d/ (D)		/k/ (K) /g/ (G)	/ʔ/
Fricative		/f/ (F) /v/ (V)	/θ/ (TH) /ð/ (DH)	/s/ (S) /z/ (Z)	/ʃ/ (SH) /ʒ/ (ZH)		/h/ (HH)
Affricate					/tʃ/ (CH) /dʒ/ (JH)		
Nasal	/m/ (M)			/n/ (N)		/ŋ/ (NG)	
Lateral Liquid				/l/ (L)			
Retroflex Liquid				/r/ (R)			
Glide/Approximant	/w/ (W)				/j/ (J)		

  - compression phase - during which a complete stop of the airflow occurs. During this time, there is no output in the acoustic waveform, and a build up of air pressure happens behind the constriction.
  - release phase - during which the articulators move to allow the compressed air to escape — this manifests as a sudden puff of air or a sudden wide-spectrum energy burst in the acoustic spectrum 2 ms in length. Frication also occurs as the articulators move open, and lasts for around 5 ms.
  - post release phase - during which there is aspiration for some of the plosives (especially word initial), which occurs before the onset of the vowel formants.

- fricatives - in which there is incomplete closure but the constriction is narrow enough to cause turbulence. They manifest in the acoustic spectrum as very wide band (nearly white) noise, sometimes strong enough to mask out the formants.

- affricates - which are each composed of a rapid coordinated sequence of a stop and a fricative. Similar to the stop, there is closure and pressure buildup corresponding to complete stoppage of the airflow, but they involve frication and turbulence at the point of release. In IPA they are transcribed with the stop and the fricative they correspond to upon release.

- liquids - which in English are the consonants /r/ and /l/, involving a complete or near complete midline closure with side branches.

- glides/approximants - which are consonants that are wide enough to almost resemble vowels in their quality.

As is typically presented, each column of the consonant chart has a pair of phonemes; the column is subdivided into the unvoiced consonant on the left, and the voiced counterpart on the right. The main distinction between voiced and unvoiced consonants is supposedly the presence or absence of vocal cord vibration. In reality, consonants in a syllable are rarely articulated without a following vowel. Since all vowels are voiced in normal speech, the distinction

thus becomes directly dependent on when the onset of voicing occurs (i.e. *voice onset time*).

The Source Filter Model and The Linear Predictive Filter

The source-filter model of speech production, depicted in Figure 1.5 [8], is similarly based on the idea of decomposing the mechanics of speech production into a source and a filter [6]; furthermore, as an engineering model, it can be used to artificially synthesize speech-like sounds [31].

The effect of the vocal folds on the airstream and the configuration of the oral cavity are respectively broken down and modeled as an energy excitation source driving an auto-regressive linear filter (a.k.a. linear predictive filter) [32, 33]. A fragment of speech can be voiced (with F0) or unvoiced. The periodic vibration from the vocal folds can be modeled at the source as a series of impulse trains driving the linear filter (i.e. glottal excitation) that represent an excitation signal fed into the resonating oral cavity [6]. When there is no voicing, air from the lungs flows through unimpeded, and it is assumed that this can be modeled as white noise driving the filter.

Recall that the oral cavity itself can be modeled as a tube with a slowly varying cross-sectional area across its length; the solution of the resonant frequencies using a discretization of the Webster equation [8] gives very good approximation to the formants in the signal. It turns out that an auto-regressive linear filter (i.e. a fed-back FIR filter) is sufficient to capture the effect of this simplified model of the oral cavity, and indeed the mathematics for obtaining the coefficients to such a filter (a.k.a. linear predictive coefficients) are directly related to the solution of the discretized Webster equation [8].

1.2.3 Perception

Speech perception is the study of how speech sounds are interpreted and recognized by the human brain. This process may be broken down into several steps:

- The acoustic waveform is mechanically transduced by the ear into nervous signals carried by the auditory nerve bundle.

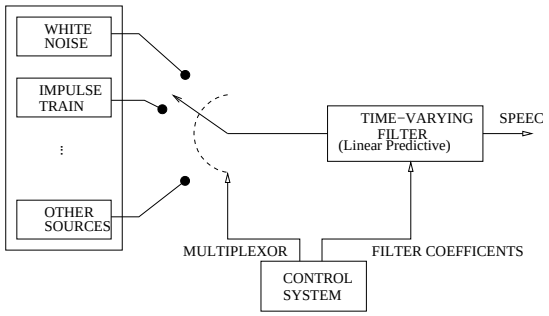


Figure 1.5: The source filter model of speech production (modified from [8]).

- The auditory nerve carries the nervous signal representing the heard sounds to the auditory cortex, where specific acoustical events may excite specific regions in the cortex in a tonotopical mapping.
- Presumably the speech events perceived by the auditory cortex may be grouped and streamed [34]. These events might be sent further upstream to processing centers for language, where the speech is eventually understood.

1.2.4 Simplified Cochlear Mechanics

Figure 1.6 shows a picture of the ear, taken from Stevens [5]. As sound impinges on the outer ear, it is filtered and acoustically amplified. The vibrational sound excites the eardrum in the middle ear, and this mechanical force is amplified through a level-like action with the incus, focusing the forces on the stapes; the acoustic energy is transmitted through the oval window into a cochlear duct (scala vestibuli) [35]. The cochlea itself is a structure resembling a snail's shell. Internally, it resembles an acoustic waveguide, separated by a thin membrane (the basilar membrane). As illustrated in Figure 1.7, the acoustics of the cochlea can be modeled by an "uncurled" version of it, that is, by a cylindrically tapering acoustic tube, divided along its length by the basilar membrane. The stapes hit the oval window, sending a traveling pressure wave down the "top" section of the tube, where it increases in amplitude until it reaches a critical position on the basilar membrane, after which it is rapidly attenuated [36]. The main effect of these intricate acoustics appears to be to cause specific sections of the basilar membrane to be responsive

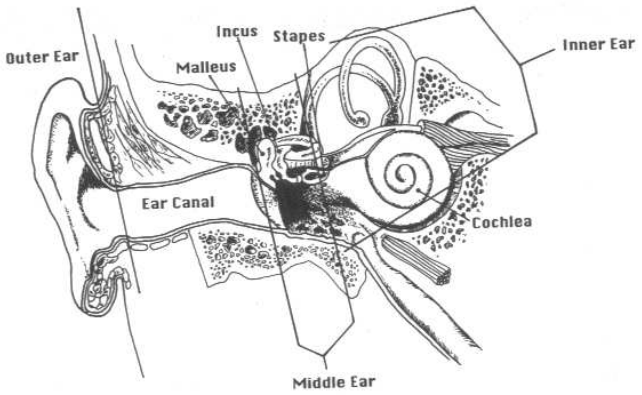


Figure 4.1 Sketch of the outer, middle, and inner ear. (From Berlin, 1994.)

Figure 1.6: Cutaway section of the ear (taken from Stevens [5]).

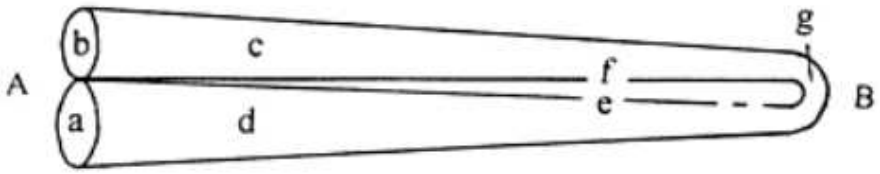


Figure 1.7: Model of the basilar membrane (taken from [38]). Here the parts are as labeled: (a) Round window, (b) Oval window connected to the stapes in the middle ear, (c) Scala vestibuli, (d) Scala tympani, (e) basilar membrane, (f) helicotrema (apical end).

to specific frequencies. This effect appears to be consistent across several species; the position along the basilar membrane with the largest magnitude of mechanical excitation for a given frequency can be modeled accurately using the Greenwood function [37] — so long as appropriate constants for the given species are used.

The transduction of mechanical energy to electrical nervous signals is accomplished by the inner hair cells [39, 40]. These cells are densely packed along the basilar membrane and are connected to nerves in the auditory nerve bundle. Increased mechanical excitation of the *stereocilia* on these cells eventually results in increased firing rate of the corresponding nerve. The cells themselves have a highly selective frequency response depending on their position, and this, combined with the natural frequency selectivity of the basilar membrane, gives extremely good time-frequency resolution of

the signal.

The transduced signal is carried by the auditory nerve from the ear to the auditory cortex, and goes through several stages of processing as it progresses through the cochlear nucleus, superior olive, inferior colliculus and medial geniculate body [5]. This processing is illustrated by Figure 1.8. Within the auditory nerve, there appears to be some sophisticated compression of the electrical signal. For instance, the firing rates of the auditory nerve have been observed to dynamically adapt to the sound source - when presented with a continuous pure tone stimulus the firing rate is high immediately after the onset of the tone, but this is quickly suppressed after the initial tone [41].

Finally, the auditory cortex combines information from both ears and perceives elementary sounds. Some researchers believe that a wide range of acoustic events have a tonotopic mapping in the auditory cortex [42, 43, 44].

Interesting Properties of the Perceptual System

The auditory and perceptual system exhibits some interesting hysteretic features and idiosyncrasies. Some interesting ones are:

- *Grouping and streaming effects.* Experiments in [34] suggest the existence of streaming and grouping effects. Psychological experiments suggest that humans are capable of both grouping, which is identifying disparate regions of energy in the spectrum as belonging to one acoustic source or event, and streaming, which is temporally chaining together disparate regions of energy as the product of a unitary physical source. These phenomena go a long way toward explaining some everyday psychological effects, such as the cocktail party effect [45], the ability of a human to "tune in" to one specific speaker amid multitudes of talking people.

- *McGurk effect.* The perception of consonants appears to be affected by cues other than those present in the acoustic source. The effect is demonstrated by the following experiment: A listener is presented with the audio recording of a bilabial CV syllable (say /ba/), and it is perceived as so. However, when the listener is simultaneously presented with a visual of the velar consonant (/ga/), the resulting consonant is perceived as somewhere in-between (/da/) [46]. The effect can be

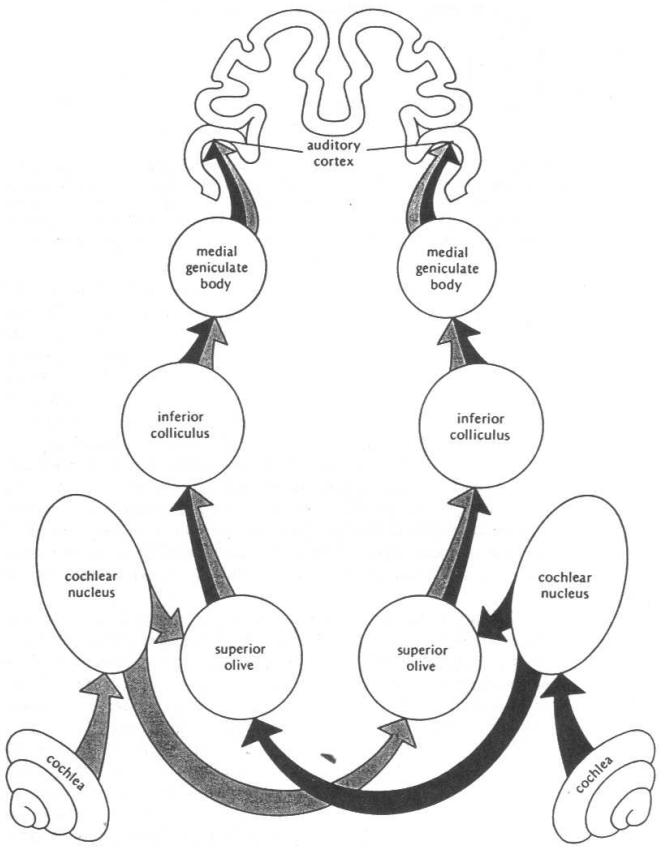


Figure 4.12 A schematic diagram of the bilateral central ascending afferent auditory pathways. (From Yost and Nielson, 1977, based on a similar diagram by Lindsay and Norman, 1972.)

Figure 1.8: Intermediate processing leading up to the auditory cortex (taken from Stevens [5]).

explained in part by the cognitive ability of the brain to extract and predict the acoustic signal from other correlated sources. It is not completely clear how much influence non-audio sources exert. In particular it is not certain if whispered speech, being possibly harder to perceive, might actually depend more heavily on non-acoustic cues.

- *Categorical perception.* The perception of speech sounds tends to fall into discrete categories which appear to correspond with the native language of the speaker [47]. Experiments with synthesized speech modifying the formant transitions to approximate intermediate transitions between /b/, /d/ and /g/ suggest that there is usually a strict boundary, whereupon the sound will suddenly be perceived to belong to the other category. The location of this boundary exhibits hysteresis; its position will be different depending on whether the sound was changing from /ba/ to /pa/ or vice versa.

- *Perceptual magnet effect* [48]. There is some evidence that when listening to speech sounds, it becomes more difficult to tell the difference between two sounds when they are similar to speech sounds of a language. Just as we can plot an acoustic space for vowels by considering the first formant on one axis and the second formant for another axis, we can similarly consider a "perceptual" space based on psychoacoustic experiments. The perceptual magnet effect seems to "warp" the perceptual space, so that discriminating sounds is easier when they are not close to phonemic sounds. This effect has been shown to be more pronounced for consonants as opposed to vowels.

At present, it is not clear at all how much of the intricacy of the human ear and the complexity of the auditory cortex needs to be emulated for accurate machine recognition. It is not clear if the level of signal detection accomplished by the human ear needs to be achieved by signal processing in machine recognition; nor is it clear that the psychoacoustic effects in speech perception have to be perfectly emulated by speech recognizers. It is well known, however, that what we have at present is insufficient, judging by the accuracy and robustness of state-of-the-art systems compared with human recognition [49]. However, it is important to keep such details in mind since speech itself is a human phenomenon, and it should be the goal of research

with machines to both recognize and misrecognize speech the way humans do. Thus, in examining what can be done with whisper, we hope that this will contribute a small piece to solving the whole puzzle of machine speech recognition.

1.2.5 Acoustics

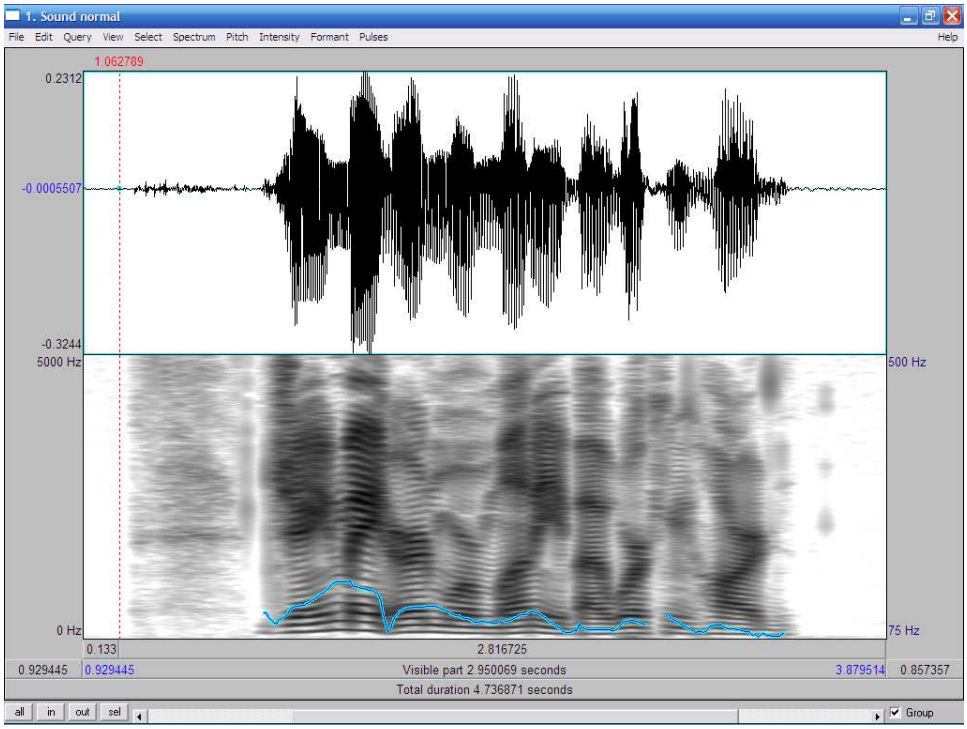
Speech manifests as a pressure wave that propagates through the air. Its study is facilitated by microphones. These measure the variations in air pressure and output a voltage signal that varies proportionally with the pressure variation. Plotting this signal against time gives us an oscillogram which we can study. The acoustic waveform gives us some information about the speech sound; large amplitudes correlate with higher volume, and sudden increase in amplitude corresponds with initial bursts of speech. The decaying amplitude indicates a general decrease in volume as we progress naturally through the utterance.

Another representation of speech is in terms of the frequency components present. We start with the short-time Fourier transform of a signal $f(t)$ computed as

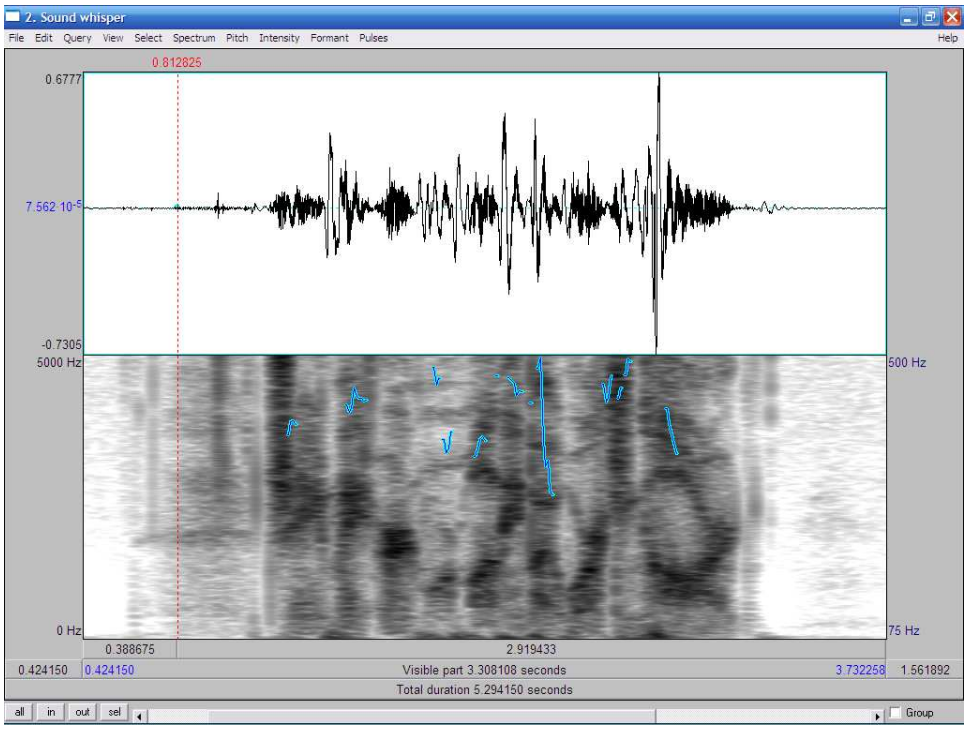
$$F(\omega, t) = \int_{\Delta t}^{\Delta_t} f(t+\theta) w(\theta) e^{-j\omega\theta} d\theta, \tag{1.2}$$

where ω is frequency in hertz, and $w(t)$ is a suitable windowing function. Plotting $\|F(\omega, t)\|^2$ with ω on the vertical scale and t on the horizontal scale gives us a spectrogram — this displays the frequency content of speech as it progresses through time. The spectrogram is able to tell us much more information about the speech signal. It has an interesting "Heisenberg" property, in that it cannot simultaneously resolve both frequency and time to a high degree of accuracy, but instead can do one or the other [50]. The length of the integration window Δt essentially controls this — a longer window is used with narrow-band spectrograms and these give better frequency resolution at the expense of time. The converse is true for wide-band spectrograms.

Examples of the acoustic waveforms and narrow-band spectrograms are shown in Figure 1.9. The pair of waveforms shown are recordings of the same adult male speaker speaking and whispering "Jane may earn more money by working hard." Much information can be gleaned from the spectrogram for normal speech. The resonances of the oral cavity now show up as dark,



(a) Normal utterance



(b) Whispered utterance

Figure 1.9: The acoustic waveform and narrowband spectrogram for both normal and whispered speech. Both recordings were of the same adult male speaker.

high energy bands (formants) during vocalic parts of the speech. Plosives show up as sudden vertical bands of energy, indicating a broad spectrum burst of energy. Many other acoustic properties of phonemes manifest in the spectrogram, allowing us to find and segment phonetic boundaries and also identify them. In fact, in many cases, a trained practitioner can deduce the phonemic identity of short segments and "hear" the utterance by examining its spectrogram [51]. The whispered versions, on the other hand, look markedly different from the phonated ones. The next chapter will go into the differences in detail.

1.3 Organization of the Dissertation

We have outlined the basics of speech communication in this chapter — it is upon this bed of knowledge that we proceed to study whispered speech. In this dissertation, we are concerned primarily with soft whisper — other types of whisper such as high effort whisper and stage whisper have different characteristics and may need to be studied separately. Specific differences between whispered and unwhispered speech will be highlighted in Chapter 2.

We introduce two new whispered speech corpora suitable for the study of speech recognition. The two corpora are the Whispered Modified Rhyme Test (wMRT) corpus designed for use in intelligibility studies, and the Whispered TIMIT (wTIMIT) corpus designed for the study and construction of large vocabulary speech recognizers. Chapter 4 describes these two corpora in detail and provides some acoustical analyses.

In Chapter 5, we consider the limit of whispering as a communication channel. A perceptual experiment and its analogous speech recognition experiment, based on the wMRT corpus, were performed. We provide an analysis of errors made by two cognitive systems, human and machine, as well as an accuracy rate for transmission of voicing in whisper at word level contexts.

In Chapter 6, we collate results from speech recognition experiments. The performance of speech recognizers at recognizing different types of speech, whispered in different accents, are considered. Standard algorithms for adapting the acoustic models, from unwhispered speech acoustic models for whispered speech and vice versa, are evaluated. We consider the problem of

building speech recognizers for whispered speech using large amounts of unwhispered speech data and limited amounts of whispered data, and propose and evaluate a new method for doing so. Finally, our conclusions are presented in Chapter 7, where implications and suggestions for future work are discussed.

CHAPTER 2

WHISPERED SPEECH

2.1 Production of Whispered Speech

The physiological production of whispered speech differs from that of normal speech primarily in the lack of phonation: vocal folds do not vibrate and the glottal aperture remains open [52]. Endoscopy studies allow us to study whisper where it differs the most — at its source — by giving us top-down images of the larynx during whispering. With whisper the larynx height is raised through actions of the stylohoid and digastric muscles [53], effectively shortening the vocal tract. This effect is also observed as a lack of phonological contrast between voiced and voiceless consonants.

2.1.1 Whisper Type and Glottal Configuration

Many researchers [54, 55] make a distinction between high effort and low effort whispering. High effort whispering, sometimes known as "forced whisper" or "stage whisper," can carry farther than low effort or soft whisper. Both types have different production parameters and acoustics. Sundberg et al. [56] go so far as to identify four different types of whispering, characterized by different airflow and glottal configuration. However, as they did not conduct perceptual tests, it is not clear if these can be really distinguished in hearing.

Monoson and Zemlin [55] used high speed laryngeal photography in conjunction with electromyography of two abdominal muscles and one strap muscle of the neck, to study differences between four registers of speech: normally phonated, soft whisper, forced whisper and breathy voice. Their study made recordings of the vowel /a/ from five subjects. Glottal apertures found during whisper include the inverted-Y and the inverted-U shapes, along

with a bow-shaped configuration.

Solomon et al. [54] observed glottal configurations of seven women and three men during low-effort and high-effort whisper. They classified vocal fold configurations into three sizes (small, medium and large) based on the shape of the glottal aperture and found two predominant shapes (toed-in, resembling an inverted-Y, and straight, resembling an inverted-U) used in the production of whisper. However, glottal configuration was inconsistent when it comes to whisper effort. This suggested to them that the production of whispered speech is motivated by the necessity to achieve salient objectives in the acoustics rather than by a need for consistent physiological production.

Similar conclusions can also be derived from Mills [53], who found that in whisper, voiced consonants are produced with a glottal configuration statistically indistinguishable from unvoiced consonants. He studied the glottal configurations of 10 speakers using a video endoscope. He also developed measures to correct the wide-angle distortion in his endoscope, and used the cuneiform tubercle as a landmark, estimating its size in pixels in order to gauge the camera to larynx distance, which was indirectly affected by larynx height. His approach allowed him to quantitatively compare the glottal aperture sizes in normal and whispered speech despite having different larynx heights. His results showed that aperture size differences between voiced and voiceless obstruents in normal speech were also observed in whispered speech. However, glottal apertures in phonated and whispered vowels were different. His work suggests that some laryngeal gestures that distinguish voiced and voiceless obstruents are preserved during whispering — he suggests that this could be a source of discriminability in the acoustics.

Work in Arabic [57] seems to give conflicting views regarding laryngeal gestures for phonological voicing in whisper. Zeroual et al. used video endoscopy to observe the larynx during productions of the Moroccan Arabic non-words /i C i C/, where C was one of the consonants /t, d, T, D, s, z, χ, ʁ, ɦ/, and /T, D/ were pharyngealized versions of /t,d/. They concluded that the glottal configuration during whisper was distinct from that during normal speech and should be considered as "whispered." They also observed that in whispered segments, the base of the epiglottis, aryepiglottic folds and the arytenoids tend to compress together. In contrast to Mills, they concluded that there were no clear laryngeal articulatory differences between consonants with contrastive voicing.

2.1.2 Transglottal Airflow and Interoral Pressure

Several different studies report higher transglottal airflow and pressure during whisper. Monoson and Zemlin [55] measured airflow during four different types of speech and found much greater airflow during whisper than in normally phonated speech. Airflow was also found mostly to be greater than soft whisper in forced whisper, and in some cases double the flow was found. These differences in production manifest in the acoustics — at higher airflow the Reynolds number of the flow exceeds the nominal value and airflow is no longer laminar but rather turbulent [29, 5]. This makes the acoustic excitation at the glottis resemble a turbulent sort of noise.

Klich measured intraoral pressure during the production of the bilabial stops under two vowel contexts /a, u/, for both phonated and whispered speech at two different volume levels [58]. Conversation level speech was produced by requesting the speakers to imagine "talking to somebody 3 feet away," and twice-conversational level was presumably double the imagined distance. He found that for phonated speech, /p/ and /b/ have different intraoral pressure, but in whispered speech they were the same. The pressure differential between intraoral and subglottal pressure gives rise to the transglottal airflow, and the two pressures depend on the acoustic impedence of the glottal constriction. Whisper, having a relatively open glottal configuration, is assumed to have roughly similar subglottal and intraoral pressures; higher transglottal airflow comes from having the lowered glottal impedence.

There are also subtle differences in both the motor control and breathing during whispering. Bonnot and Chevrie-Muller [59] made electromyograph measurements of the activity of three muscles: the *orbis oris inferior* surrounding the lips, the *anterior digastric* on the underside of the jaw, and the *levator veli palatini* which elevates the soft palate. By aligning these signals with the start of the acoustic waveform, they found that during whispering speakers tended to have longer anticipatory signalling than in phonated speech.

2.1.3 Breath Control in Whisper

Breath control differs from normal to whispered speech. In [60], the respiratory function during whispering was investigated for 10 healthy subjects.

Subjects were made to whisper and speak a single paragraph. Syllables were grouped together for each draw of breath, and measurements were made for the various lung capacities during whispering and speaking. From these, the amount of air expended during whispering was calculated and compared with normal speech. They found that each draw of breath had roughly the same amount of air during the production of both whispered and normal speech. They also found that whispered speech production was slower, more air was expended per syllable, and fewer syllables could be spoken. These findings are corroborated by Schwartz [61], who found that more air was used during whispering. He also suggested that in order to conserve air, gestures which conserve airflow such as stop closures are prolonged, leading to an overall lengthening of whispered syllables [62].

2.1.4 Articulator Movement in Whisper

Higashikawa investigated the differences between lip movements for bilabial plosives during phonated and whispered speech production [63]. Speakers were asked to produce CV syllables with /p/ and /b/ in them set within a sentence context. To track lip movements, reflective markers were placed on the lips of the seven subjects and these were videotaped and later automatically tracked. The authors found significantly faster lip opening when whispering /b/ compared with whispering /p/ or normally speaking /b/. This work is suggestive in that one might now suspect that hyper-articulation exists for the other articulators of speech; however, beyond the literature listed I am not aware of any evidence that different or exaggerated articulatory movements occur during whispering.

2.2 Acoustics of Whispered Speech

Acoustically, whispered speech is very different from non-whispered speech. We once again refer to Figure 1.9 for comparisons. A cursory examination of the waveforms show that the whispered version looks nothing at all like the normal speech waveform. The sudden peaks in the waveform correspond to plosive bursts, which have a tendency be much stronger (relatively) in whispered speech. The spectrograms indicate that no voicing is present at

all in the whispered speech, as there are no visible horizontal striations that typically signal voicing. Although it is not apparent, the properties of increased spectral tilt observed by researchers in speech perception seem to be present. The obvious remaining indicators of the message for whispered speech appear to be largely the formant energies.

2.2.1 Reduced Spectral Tilt in Whisper

Without glottal fold vibration, voicing information appears to be lost. There is no viable concept of voice onset time, which is otherwise typically cited in the literature to be a feature that distinguishes between contrastive voiced and voiceless phonemes. The glottal source behaves roughly like a noise source, and the spectral quality of most phonemes is changed. A general observation is that whispered speech has a lower, but flatter, power spectral density (i.e. less spectral tilt) compared with phonated speech [64].

2.2.2 Shift in Formant Frequencies in Whisper

Many studies have also found a change in frequency of the formants when whispering. Kallail and Emanuel [65] recorded isolated phonated and whispered vowels /i,u,æ,a,ʌ/ from 15 male adult speakers. Spectrographic measurements showed systematic increase in the first three formants. They also found that `F1` was modified far more than `F2` or `F3`; this was attributed to the change in glottal vibration effectively shortening the overall length of the vocal tract, thus substantially modifying `F1`. They assumed that the position of the tongue tip remained unchanged during whisper, and thus `F2`, which tends to be associated with resonances due to the front oral cavity, remains less changed. Further results obtained by Kallail and Emanuel for English [66], along with those obtained by Slobodan for Serbian [67, 68], and Itoh for Japanese [69] tend to be similar.

Slobodan's measurements of the formant frequencies [68] for the vowels /i,e,a,o,u/ for five male and five female speakers of Serbian showed a rise in `F1` in all vowels except /u/, a rise in `F2` for all except /u/ and /e/ for males, and relatively unsystematic changes in `F3` and `F4`. One drawback of his data is that mean values for the formant locations were taken, instead of per-speaker

difference, which would get rid of variability due to speaker. Also, he found that the formant bandwidths in whispered speech were systematically larger than in phonated speech.

Itoh et al. [69] recorded whispered and phonated speech from 69 male and 49 female speakers. Speakers read from a collection of 60 sentences, 50 of which were phonetically balanced for Japanese. Measurement of the formant frequencies showed a general trend to increase formant frequency for the lower frequency formants (i.e. `F1` and `F2`), but they do not make any observation on the higher frequency formants.

2.2.3 Intensity of Whisper

Whisper tends to be soft; signals captured with conventional microphones have a low signal-to-noise ratio. Furthermore, since whispered speech is quite similar to spectrally shaped noise this hampers algorithms which attempt to denoise or improve SNR. Despite all of this, the intelligibility of whispered speech does not appear to fall far below that of normally spoken speech [70]. Furthermore, there is even surprising evidence that certain information (e.g phonemic voicing distinction, or emotion [71]) not expected to be conveyed well, actually is.

2.3 Perception of Whispered Speech

There have been a number of studies involving the perception of whisper. Since voicing is absent, acoustic pitch is non-existent: one does not expect pitch and pitch-related information to be conveyed. However, there is much evidence from the literature to suggest the contrary.

2.3.1 Auditory Nerve Representation of Whisper

How the auditory system represents whispered speech has also been a subject of research. Evidence from Stevens and Wickesberg [72] suggests that voicing distinction is made early in the auditory system. In that work, audio recordings of the syllables /ta/ and /da/, phonated and whispered, were presented to anesthetized chinchillas, and auditory nerve recordings taken.

Global time-averaged peri-stimulus time diagrams for both syllables were found to be distinct, suggesting a different neural encoding for voiced versus unvoiced phonemes in whisper. Furthermore, they found an exaggerated double onset in the neural signal in response to /da/ aligned with the plosive burst and vowel onset, which was not found in the response to /ta/. This suggests that a sudden burst in energy a second time after the plosive burst is a discriminating characteristic of the voiced phoneme. Thus, there is already evidence that the inner ear and the process of encoding in the auditory nerve treat voiced and unvoiced phonemes differently in whispered speech. The following perceptual results therefore do not seem very surprising.

2.3.2 Perception of Vowels

Kallail and Emanuel conducted perceptual experiments to determine the identifiability of the vowels /i,æ,ʌ,a,u/ when whispered [65]. Spoken and whispered tokens were collected from 20 female adult speakers, and presented to 2 panels of 11 listeners. Listeners, who were all graduate students in linguistics, were allowed to choose from the vowels and glides /ɪ,ɛ,æ,ʌ,ɑ,ʊ,u/ as responses. He found that listeners could identify the correct vowel 85% of the time in phonated speech, and 63% of the time in whispered speech. Analysis of the confusion matrices indicates that errors tended to occur with vowels close by in the vowel space. A second experiment using essentially the same type of test utterances but with 15 male speakers gave the same result — that vowel identity was less accurately conveyed in whisper than in phonated speech, but still reasonably well conveyed. Although the authors claim that whispered vowels "lack acoustic features important to vowel identification," this does not appear to be the case since in many cases formant structure is well preserved.

A similar experiment was conducted by Tartter [73]. She recorded spoken and whispered versions of the vowels /i, I, ɛ, æ, a, ɔ, u , ʌ, U, ɝ/ within the consonant context `[h _ d]`, from three male and three female speakers. Twelve listening subjects were first familiarized with the vowels by having the experimenter review and produce voiced versions of them. The subjects were divided into two groups; the first group recieved additional instruction repeating the live-voice demonstration, and was also familiarized with speak-

ers in the test data. Each group was administered a perceptual test with two sets of stimuli, one whispered and one phonated. Overall results showed good identifiability of each vowel, ranging from 80% to 99% for phonated tokens and 72% to 99% for whispered vowels. Per-vowel accuracies did not drop by more than 20%, indicating a high rate of identifiability of vowels in whispered speech.

2.3.3 Perception of Phonemic Voicing in Obstruents

There is consistent evidence that phonemic voicing is conveyed in whisper despite the lack of glottal vibration. Some of the earliest work by Dannenbring [74] confirms this. In this work, the author recorded his own whispered CV tokens from the 12 consonant contexts /b,p,d,t,g,k,z,s,v,f,ð,θ/ and the three corner vowels contexts /i,a,u/. Listeners were presented with these CV tokens and given a choice of the correct consonant or the consonant with opposing voicing distinction. The 12 listeners who participated were also asked to rate their confidence of each judgment on a scale of 1 to 7. This effectively put each judgment on a 14-point scale, spanning from confident unvoiced judgments on one end to confident voiced judgments on the other end, from which the rank-based D statistic was computed [75]. His results showed that subjects were both confident and correctly judged most voicing-distinct opposed pairs. Good judgments were obtained for the plosives in all vowel contexts, but poorer judgments were obtained for the fricatives and affricates. In summary, listeners were found to be able to discriminate between whispered voicing distinct phonemes with confidence.

One principal study by Tartter [76] involved the perceptual study of whispered consonant-vowel syllables. We can think of this study as analogous to Miller and Nicely's classic experiment, but for whispered speech. Her stimuli consisted of so called "nonsense" syllables with a consonant and following vowel (CV), produced by a single male and a single female talker. The 18 consonants used were /b,d,g,p,t,k,m,n,l,w,y,v,f,z,s,ʃ,ʒ/. Notice that the affricates were omitted from this experiment. Just as in Miller and Nicely [77], the vowel /a/ was used. Six listeners involved in the study were asked to identify the consonant in the first experiment, and the speaker sex in the second experiment. However, results from their gender identification experi-

ment cannot be generalized due to the small number of speakers, as listeners could be distinguishing them based on individual characteristics rather than properties pertaining to gender.

Confusion matrices were constructed for the consonant identification task by accumulating the responses from the listeners: for each response the stimulus gives the row and response gives the column of the confusion matrix, for which a count is accumulated. Articulation indices were computed from these matrices using Miller and Nicely's formula. Tartter assessed how much linguistic information was transmitted by collating the matrices into smaller ones based on the desired feature, and computing *information transmission* for each feature

$$T(x) = \sum_{i,j} \frac{n_{i,j}}{n} (\frac{log_2 n_i/n \times n_j/n}{n_i j/n}), \tag{2.1}$$

where $n_{i,j}$ is the entry in the collated confusion matrix, n is the number of categories, and n_i and n_j are row and column marginal sums. This approach can be viewed as computing the joint entropy of $P(\hat{\theta}, \theta)$, where θ corresponds to the produced category and $\hat{\theta}$ corresponds to the perceived category, for which $n_{i,j}/n$ are estimators. Her results indicated a 64% accuracy rate for identification, 0.85 bits per stimulus for transmission of voicing, 0.94 for place, and 0.61 for manner. It is, however, not clear what exactly contributes to this transmission.

One interesting result from the confusion matrices is that [voiced] $\rightarrow$ [unvoiced] errors tend to happen far more often than the reverse. One way to interpret this is that in whisper voiced phonemes tend to more closely resemble their unvoiced counterparts.

For the /b/ and /p/ phonemes, Munro [78] has suggested that the difference in relative intensity between the consonant and successive vowel is a contributor to voicing discrimination. In his first experiment, he took a total of 32 CVs with the /p,b/ consonant and /æ,ɛ,i,u/ vowel contexts from two male speakers and examined their oscillograms. He defines two statistics: the rise times t50 and t75, representing the time it takes from the onset of the burst to respectively reach 50% and 75% amplitude of the mean amplitude of the following vowel.

His measurements seem to indicate that /b/ tokens have slower rise time than /p/. Perceptual experiments were conducted on six female and two male listeners. Overall, an accuracy of 63% was obtained for /p,b/ discrimination.

Unfortunately, the number of tokens in the experiment was small and thus the statistics unconvincing. Furthermore, tokens that were more prototypical of this feature — that is /p/'s with faster rise times, and /b/'s with slower rise times — did not appear to be identified with greater accuracy.

2.3.4 Perception of Pitch and Tone in Whisper

Several studies have concluded that listeners can perceive pitch in whisper, although exactly how is probably still an area of contention. Thomas [79] asked listeners to listen to whispered vowels and specify their pitch by setting the same pitch on a pure tone oscillator. He found that the pitch set by the test subjects often corresponded with the location of the second formant located with acoustic analysis.

Higashikawa et al. studied the perception of whispered pitch [80]. Six male and six female native speakers of Japanese were asked to whisper the vowel /a/ in three pitches: ordinary, high and low. The subsequent listening test with five otolaryngologists found accurate identification for 11 of the speakers. The first three formants of accurately identified vowels were examined, and pitch was found to systematically correlate with the frequencies of the first three formants. In another of their studies [81], they found evidence to suggest that whisper pitch perception occurs in a more complex way: that it is influenced by simultaneous changes in `F1` and `F2`.

Cheung [82] investigated whisper pitch in Cantonese, a language known to be highly tonal with nine tones. The author recorded stimuli for 4 pronunciations with 6 tones from 3 male and 3 female subjects and conducted a tone identification experiment with 12 listeners. His results found tone identification for some of the tones to be above chance, at an overall identification rate of 22%.

Many other languages also use pitch phonemically: sounds that are essentially the same except for the pitch contour mean different things. For instance, Thai is known to have 5 different tones [83]; Mandarin has 5 tones [84]; East Norwegian has pitch contours [85]. Abramson presented his subjects with sets of whispered Thai words [83], each containing 4 to 5 distinct words which differed only in tone. His first result involving monosyallabic words was inconclusive; identification rates hovered around chance. A sec-

ond experiment involving groups of words with different tone but whispered in the same sentence contexts showed markedly improved identification. His results seem to suggest that information pertaining to tonality is distributed in longer contexts.

A similar experiment in East Norwegian involving sentence level contexts was conducted by Nicholson and Teig [85]. The authors devised a series of sentence pairs which were identical in the front up to a position where either word from a pair of similar words with different tone was found. They played back spoken, whispered and resynthesized versions of these utterances up to and including the tonal word, and asked listeners to choose one of two options to complete the sentence. Their results showed that listeners were able to identify the tone in whispered speech up to a 61% accuracy. These figures were found to be above chance, and suggest that tone information is conveyed in whispered speech.

Gao [84] conducted a detailed study of tones in whispered Mandarin. In Mandarin, tone is conveyed by a number of factors - notably through pitch contour, though other cues such as amplitude contour exist. Acoustic data were collected from 2 male and 2 female speakers. Her stimuli consisted of the syllables /ba/, /fa/ and /ma/ in all four tones in both isolated and a paragraph level context. For the longer contexts, a pair of speakers were made to enact a small conversation in which the syllables were embedded. Acoustic measurements found a longer syllabic duration in whisper. In some cases, especially for females, a more exaggerated amplitude contour was found. Perceptual experiments with 10 female listeners found over 90% accurate identification for spoken tones and 60.1% in whisper. The author suggests that the most important contributors to perception are the whisper's "special maneuvers" to exaggerate acoustic properties that correlate with tone, amplitude contour, and semantic context.

When taken together, these studies are inconclusive and appear to suggest that the acoustic cues for tone vary from language to language. This also must be the case when whispering. It is clear, however, that in some languages tonality is conveyed.

2.3.5 Perception of Speaker Identity and Gender in Whisper

Voice is an intrinsic part of identity, and they are "unique like personal faces" [86]. One key component of identity is speaker gender. Tartter's second experiment in [73] dealt with the speaker identification in whisper. The 12 subjects involved in the experiment were familiarized with the phonated versions of the utterances, and then asked to identify the speakers for the whispered versions. The overall accuracy for each speaker ranged from 46.2% to 62.5%, well above chance at 33%. Some listeners were highly competent, obtaining as high as 96.3% accuracy. Her results suggest that certain acoustic cues pertaining to speaker identity can carry across whisper, and she suggests that speaker syllable duration is one of these.

Lass et al. [87] recorded whispered isolated vowels from 10 male and 10 female speakers, and conducted a perceptual experiment with 15 listeners. He found a 75% accuracy rate for whispered vowels and 95% accuracy for phonated vowels. His results clearly confirm that gender information is carried in whisper. Whether other facets of speaker identity carry through is not clear, and there is room here for further study.

2.3.6 Perception of Emotion in Whisper

There is some evidence that emotional cues can be found in whisper. In [71], Tartter recorded CV utterances from three male and three female speakers of North American English. Three types of speech were produced: in the first type, speakers were told to physically smile but "try not to sound happy." The second type was similar but done for frowning. In the third type, speakers produced whispered speech while physically smiling, but tried not to sound happy. Results from listening tests involving six listeners showed that they were able to detect physical frowning in normal and whispered speech. They could detect physical smiling in normal speech, but did not seem to be able to detect physical smiling in whispered speech.

2.4 Applications of Speech Technology to Whispered Speech

We now turn to some more recent work involving speech technology, of which there is less literature. Most studies to date have involved the application of automatic speech recogntion and model adaptation from phonated speech to whisper. These methods have generally used the most basic techniques. One particular interesting application is the morphing of whispered speech to voiced speech. There are several approaches in the past that deal with this.

2.4.1 Detection of Whispered Speech

As a precursor to processing, whispered speech has to be segmented from non-whispered speech. Carlin et al. [88] describe an algorithm for detecting whispering in the midst of phonated speech. They employ two features based on different spectral tilt in whisper and the lack of voicing in order to distinguish whisper from normal speech. The first feature is a ratio of high frequency energy to low frequency energy, 2.5 kHz being the cutoff for the different frequency bands; this feature can capture the reduced spectral tilt in whisper. The second uses LPC analysis and applies an inverse filter in order to get at the residual signal. If we accept that LPC is a good enough model for the transfer function of the vocal tract, we are left with the residual that corresponds to the glottal excitation. Modified autocorrelation was performed on a cubed version of the residual signal, by computing Pearson's linear correlation coefficients between data in the first and second halves of the input frame. After this, peak picking was applied to extract the maximum autocorrelation — whispered frames of data would have small amplitude, but voiced frames would have large amplitude. Finally clustering was applied to classify whispered and phonated speech frames. Their approach was found to be able to correctly detect whispering 97.5% of the time.

2.4.2 Enhancement and Recognition of Whispered Speech

The most recent and significant work on the recognition of whispered speech is the doctoral dissertation of Robert W. Morris [70]. In this work, a number of separate studies were conducted, with the intent of improving an existing technique of phonating whispered speech to produce normal sounding speech. Although the dissertation covers algorithms and estimation techniques for speech enhancement, noise and removal, I shall only deal with two sets of studies most relevant to this dissertation.

The dissertation relies heavily on the Diagnostic Rhyme Test (DRT) [89] as a means of discerning which part of the acoustic channel encodes distinctive features for speech. This DRT proceeds by presenting a listener with audio recordings, which have one of two possible words embedded in a carrier sentence. The chosen words are minimal pairs, differing by only one type of distinctive articulatory feature, be it voicing, nasality, sustention, sibilation, graveness or compactness. The listener is faced with a binary decision; the raw scores for DRT are the normalized difference between the correct and wrong responses to the test. This same test can also be administered to an artificial system such as an automatic speech recognizer.

The first set of studies dealt with the intelligibility of normal and whispered speech under differing noise and speech coding conditions. To do this, the author recorded a small corpus of whispered and normal speech, uttered by three male and three female speakers, under three types of noise conditions (quiet, office and street cafe). The utterances were selected from a set of 15 phrases, and 232 isolated words, with the intent of conducting the DRT and Diagnostic Acceptability Measure (DAM) respectively. The utterances were recoded using three different codecs (CVSD, MELP 2400, MELP 2400 MPP), and the DRT administered to eight listening subjects. Their results under quiet recording conditions showed that the majority of confusions for whispered speech are associated with the voicing feature. However, when speech coding was applied, the intelligibility of unvoiced whispered words took a severe hit, indicating a possible inadequecy of the investigated algorithms (especially MELP) at encoding whispered speech. Under noisy environmental conditions, there is a reduction in intelligibility of the voicing distinction, but these can be ameliorated by noise-enhancement algorithms.

One particularly stunning result from this is that applying MELP does not

degrade the voicing information of whispered speech, when the whispering is done in a noisy environment. There are several possible theories as to why this is the case; perhaps in a noisy environment, people whisper in a different way, emphasizing the acoustic cues which just so happen to be carried across through the encoded channel. If this is indeed true, then whispering under quiet versus noisy conditions might as well be considered as two distinct types of acoustic variability that the universal speech decoder needs to compensate for. The results also indicate the importance of applying well designed noise-enhancement algorithms as a preprocessing filter to the speech recognition system.

The second set of studies examine the efficacy of traditional, well established automatic speech recognition (ASR) techniques at recognizing whisper, once again using the DRT as a test framework. In all of the experiments, the commercial HMM-based Fast-Talk tool was used. For normal, full voiced speech, the machine recognizer performed worse than the human listeners, but was still able to distinguish the minimal pairs corresponding to all six distinctive features. This is not the case for whispered speech: the test scores suggest that the machine recognizer is making a decision at chance levels for two types of cepstral features. Furthermore, no amount of adaptation appears to improve this problem of voicing distinction.

By carefully interpreting the data from the intelligibility tests in conjunction with the automatic recognition tests, we can obtain key insights into what needs to be done to improve the state of the art. Although the experiments themselves are quite thorough, there are a few gaps which need to be filled. Most importantly, the question of how voicing distinction is carried across the communication channel remains unsolved: it is clear that the ASR system cannot distinguish voiced obstruents from their unvoiced counterparts for whispered speech, but does the problem lie with the feature parameterization, or is it an issue with the pattern recognition algorithm? It is unfortunate that the human-listening tests neglected to include resynthesized waveforms using MFCC features, as this would give us a conclusive demonstration of the inadequacy of MFCC for recognizing whispered speech.

Although Morris' dissertation provides some very interesting ideas, there appear to be some drawbacks with the methodology. First, the speech corpus is too small and too varied in terms of the noise and environmental conditions: statistical variances are possibly not sufficiently ameliorated by the

lack of data. Second, the choice of a proprietary, closed-source tool, employing a relatively untested jump Markov linear system (JMLS) as the pattern recognition backend (as opposed to the HMM toolkit - a more well established research tool), casts doubt on the strength of the conclusions that can be drawn from the DRT score data. Furthermore, the author himself admits that JMLS training requires that a large number of parameters (exponential with regard to the length of the signal) be estimated, further exacerbating the problem posed by data insufficiency. Clearly, these issues must be addressed in a follow-up experiment, in order to demonstrate beyond doubt that there is a problem with the current acoustic feature extraction, and conclude that research in voicing distinction is the missing gap in our understanding.

2.4.3 Automatic Recognition of Whispered Speech

In 2002, researchers at the University of Nagoya [69] collected an audio-visual corpus of whispered speech. Their intended application was to develop a speech recognizer specifically capable of handling whispering on cellphones. A parallel corpus of normal and whispered utterances were collected for Japanese. A total of 68 male and 55 female speakers read phonetically balanced sentences, and audio and visual recordings were made under two types of recording environments: close talking microphone (CTM), and telephone handset (TH). For the CTM setup, recordings were made in a soundproof room at high fidelity, whereas with the TH setup, recordings were made in both a soundproof and less quiet environment (their computer room); the codec (32 kbps ADPCM, G.726) for the personal handphone system (PHS) was applied.

At least three mini-studies relevant to our work were conducted. First, their examination of the speech spectra indicated, on average, an upward shift in both the first and second formant frequencies. The effect is more pronounced and consistent with the first formant, but is only marginal with the second formant, especially in the case of the vowels /i/, /u/ and /e/ which sit at the extremities of the vowel chart.

Next, the authors examined the averaged spectra of each phoneme class, and computed the cepstral distances between the averaged spectra for each specific phoneme. Their phone segmentation and alignment was obtained

automatically using dynamic time warping with the Itakura distance measure. The results, tabulated for all the phonemes, indicate a drop in energy near the low frequency band; this is consistent with the idea that voicing is absent, and thus pitch energy is lower. However this effect is also present for the unvoiced consonants.

Measurements of the averaged cepstral distance for each phone class show that the vowels, glides and nasals differ the most going from normal to whispered speech, followed by the voiced plosives and alveolar fricatives, then finally unvoiced plosives, affricates and other fricatives. The next set of studies deal with automatic recognition, both at the phoneme level, and with a full word recognizer. Using HMM-based techniques, the authors train whispered and normal speech acoustic models from the data, and evaluate syllable and word recognition accuracy for both types of acoustic models on both types of speech. The outstanding result from these experiments is that the whispered speech model performs almost as well on either type of speech. The authors next use MLLR adaptation to improve the ability of their normal speech model to recognize whispered speech. With merely 10 utterances of a target speaker (i.e. a closed test), the accuracy of the normal speech model improves to within range of a whispered speech model. Adaptation with a development set on an open test shows significant improvement, but with room for further performance gains. Thus, a reasonable conclusion to draw is that MLLR itself is a robust enough technique to allow a more readily available normal speech model to recognize whispered speech.

Error studies of the confusions made by the whispered phoneme recognizer indicate that a relatively poor model was trained. What is more interesting is their result with full word cross-mode recognition. In this setup, the same methodology was used for training acoustic models of whispered and normal speech. Next, these models were set up with a word grammar and used to perform recognition on both types of speech. (I.e. a whispered speech model was used to recognize both whispered and normal speech, and normal speech model to recognize both whispered and normal speech.) Their results are nothing short of stunning: the whispered speech model worked as well as the normal speech model for recognizing normal speech. There does not appear to be a clear reason for this result, and in fact our experiments do not confirm it. It is crucial to perform a more thorough investigation of this matter.

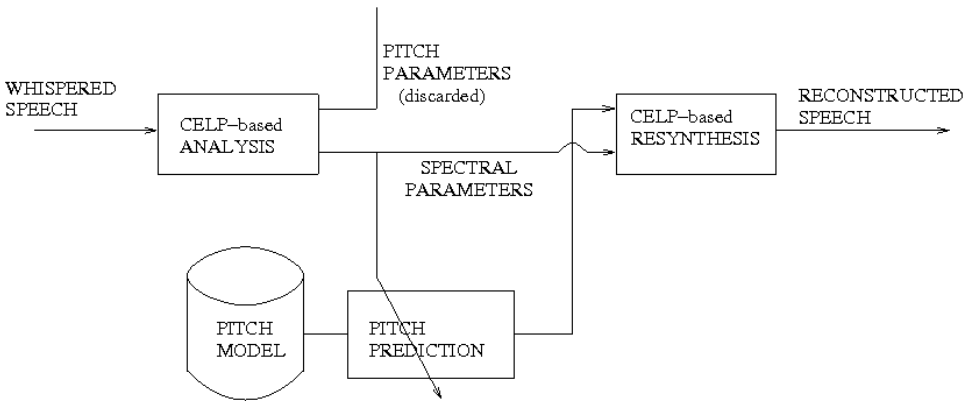


Figure 2.1: Basic framework for reconstructing voice.

2.4.4 Resynthesis of Phonated Speech from Whisper

One interesting application is to attempt to reconstruct phonated speech from whisper. This is motivated in part by the need to improve the quality of life for post-laryngectomy patients, allowing them to "speak" with a normal voice in spite of their dysfunctional larynx. The majority of approaches employ methods from speech coding — a basic framework is illustrated in Figure 2.1. The basic idea is to assume that the synthesis parameters associated with voice are missing, and to find a way to reintroduce them.

Morris [70] provides an algorithm based on this approach using mixed-excitation linear prediction (MELP). He explicitly models voice parameters using a JMLS, and uses this to reintroduce pitch. Additional refinements include a Wiener filter to remove breath noise from the whispered speech, and also a frequency warping algorithm to compensate for the different formant locations in voiced and whispered speech. Multiple utterances were resynthesized, applying a combination of true (original voiced speech) parameters and parameters generated from his algorithm. His results indicated that using modified spectral parameters hurt DMOS scores more than using synthetic pitch parameters. This suggests that a more intricate algorithm for handling formant shift is needed.

A more recent work by Sharifzadeh [90] explores the same concept but using code-excited linear prediction (CELP) [91]. Their approach works by reintroducing pitch, which is estimated using a dynamical system. On examination, the reconstructed pitch contours seemed to mimic naturally occurring pitch. However, as no subjective tests were undertaken we cannot fully

evaluate the quality of their algorithm.

CHAPTER 3

AUTOMATIC SPEECH RECOGNITION

3.1 Overview of LVCSR

Large vocabulary continuous speech recognition (LVCSR) is an ongoing and active field of research, as it has been for the past 30 years [92, 93]. The most successful methods to date are based on the hidden Markov model, which was introduced to speech recognition by Baker [94, 95] and Jelinek [96]. Many systems today employ this approach (e.g. HTK [97], Sphinx [98, 99], Julius [100, 101], Decipher [102]). This is not to say that other approaches do not exist; for instance, segment-based methods [103, 104, 105] and finite-state transducers [106, 107] are also in use. Arguably, methods like dynamic Bayesian networks are a generalization of the HMM approach, and are quite similar to it [108]. Even more techniques, such as point-process models, are still being proposed today [109]. A thorough categorization and review of all these techniques would be beyond the scope of this dissertation; instead we will focus on the algorithms that we applied in our experiments.

In this section we will review the HMM-based approach [94, 110, 111] and its associated techniques. This approach to LVCSR is probabilistic in nature and makes some assumptions about the nature of speech that at times have been criticized by speech scientists [112]. The objective of probabilistic speech recognition is to find, given a sampled acoustic signal, the best matching hypothesis of what words were actually said [113]. In other words, we want to find the optimum sentence

$$S^* = \underset{S}{\operatorname{argmax}}\, P(S|O), \tag{3.1}$$

given a sequence of observations $O = (o_1, ...o_t)$ on the acoustic speech signal. The sentence $S = (w_1, ...w_m)$ is a hypothesis built up of an arbitrary m

number of words $w_1, ..w_m$ chosen from a fixed vocabulary W. In statistical speech recognition, the maximization is performed with a given model λ of spoken language. If we treat the probabilistic model as a generative model (e.g. a hidden Markov model), then the probability distribution is dependent on the given model, and Equation 3.1 becomes

$$S^* = \underset{S}{\operatorname{argmax}} P_\lambda(S|O) \tag{3.2}$$

Now, λ itself could represent any reasonable statistical model of spoken language, but a completely unstructured model for spoken sentences would require so many parameters to be estimated that training it becomes intractable [114]. Spoken language is hierarchical in nature [115]: sentences are made of words, which are made of sub-word units such as phonemes, which in turn are made of landmarks; this hierarchy can be exploited to factorize the number of possible variations and thus parameters required at each level. Most commonly, we consider the full model as a product of parameters modeling high level language (i.e. typically the language model λ_{lm} that models word probabilities) and the parameters modeling acoustic subunits (i.e. phonemes or landmarks, using an acoustic model λ_{ac} [116]). In a practical system, a pronunciation dictionary, not necessary probabilistic, will also be needed (i.e. a pronunciation model λ_{pron} [117]). This breaks down the parameters that need to be estimated for spoken language as

$$\lambda = \lambda_{lm} \times \lambda_{pron} \times \lambda_{ac}. \tag{3.3}$$

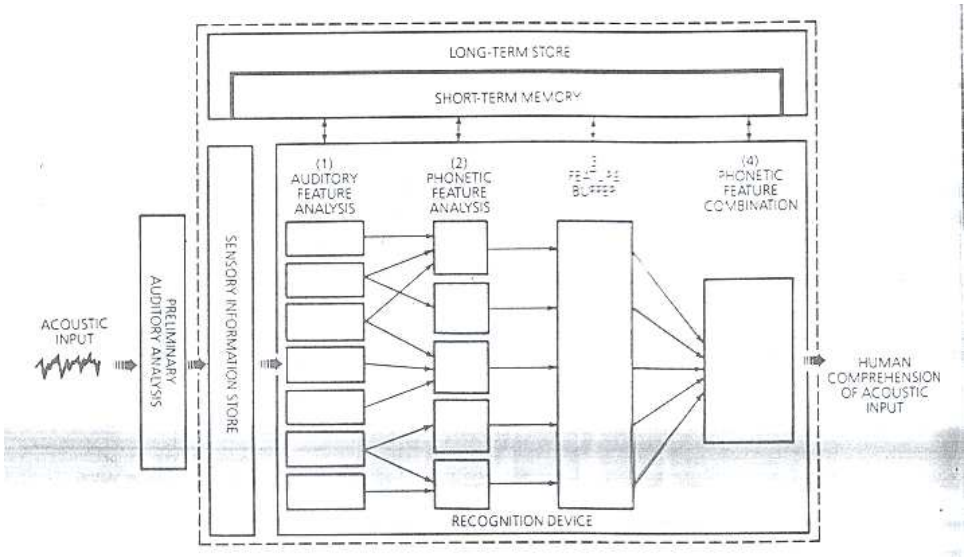
Applying conditional probability, Equation 3.2 becomes

$$P_\lambda(S|O) = P_{\lambda_{lm}}(S|W) P_{\lambda_{pron}}(W|\Phi) P_{\lambda_{ac}}(\Phi|O), \tag{3.4}$$

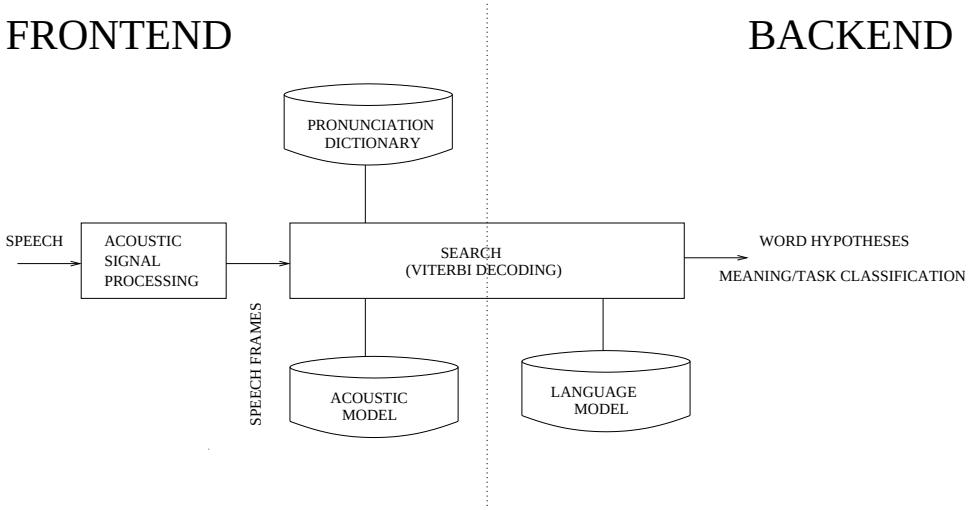
where Φ represents a sequence of sub-sentence units (usually phonemes), and λ_{ac} and λ_{lm} represent statistically estimable parameters that make up the *acoustic model* and the *language model* respectively.

Almost any state-of-the-art system (shown in Figure 3.1(b)) can be decoupled along these lines, and will have two halves: the *front end* consisting of signal processing and acoustic pattern recognition (with *acoustic model* λ_{ac} and pronunciation model λ_{pron}), and the *back end* that enforces linguistic

constraints with the language model λ_{lm}. In practice, systems perform decoding simultaneously, so this distinction between pattern recognition and enforcing linguistic constraints is somewhat blurred.



(a) Model of human cognitive perception of speech. (from David Pisoni)



(b) Architecture of a generic large vocabulary automatic speech recognizer.

Figure 3.1: Speech recognition by humans versus machine.

This architecture has sub-components which are directly analogous to the functional parts of human cognitive speech processing. A suggested model by Pisoni is shown in Figure 3.1(a), where we can see some parallels: for instance, the signal processing is analogous to the human ear, the acoustic pattern recognition to what is done in the cortex, and the enforcement of

linguistic grammar to what is done by our higher-level modules of language processing. As speech itself is native to natural human-to-human communication, there is no reason to believe that an alien architecture could outperform the established setup for this task. Nonetheless, with the possible exception of speaker [118] or language identification [119], experiments in discrimination and classification tasks demonstrate that the best performance of artificial sub-components of these systems does not match that of humans at most common cognitive tasks [49] — the best recognizers are neither as accurate or robust as humans.

3.1.1 Frame Synchronous Speech Recognition

The HMM-based technique is *frame synchronous* in nature [120]. That is, it is assumed that the incoming speech can be analyzed in terms of frames – these are essentially vectors that represent the signal sampled at regular intervals of time. They are usually produced by analyzing the signal at regular windowed intervals, using a technique such as the short-time Fourier transform (STFT), although more elaborate time-frequency techniques involving the Wigner distribution [50] or wavelet analysis [121] can be used. The recognizer briefly comprises the following stages, whose detailed workings are described in the sections following.

- Initial Signal Processing - In this stage the acoustic wave-form is converted into a frame. The most common method is to employ a filter-bank and cepstral computation. Some noise cleaning may be performed before filter-bank analysis, and additional post processing can be used to normalize channel effects.

- Acoustic Pattern Recognition - In this stage, the feature vectors are recognized and categorized into sub-word units. The most common acoustic target used are words, although other subword units like syllables or phonemes or features such as landmarks [122, 123] could also be used. The recognizer is influenced by trained acoustic models, language bigrams and the pronunciation dictionary as shown in Figure 3.1(b). The end result gives us the best or a list of n-best transcriptions, which could then be decomposed into the best fitting sequence of phones.

The acoustic model itself could be modeled using any probabilistic graphical model (e.g. the hidden Markov model (HMM) [110], maximum entropy Markov models (MEMM) [124, 125], dynamic Bayesian networks (DBN) [108, 126]). An embedded word graph is constructed by embedding phone-level graphical models into a word-level graph (see Figure 3.2), each node in the final graph corresponding to a specific model state. Using Viterbi decoding, nodes in the model are "activated" at varying likelihoods - the transitions out of the dynamic state of each graph node usually represent the detection of an acoustic event. After pruning away unlikely hypotheses, the most likely word sequences can be compactly represented in a lattice.

- Enforcement of Linguistic Constraints - In this stage, the high level linguistic knowledge is imposed on the lattices or phone strings to obtain actual word hypotheses. The knowledge used here typically corresponds roughly to the grammatical components of *syntax*, *pragmatics* and *semantics*. In the case when the job of the recognizer is limited [127], the constraints can be imposed by a task-dependent regular grammar. In large vocabulary recognition, where the user is allowed to say anything he or she wants, a more permissive model is required. The most common of these are n-gram models [128]. Sometimes, it is not necessary to recognize whole words for an intended application; an example of this is in language identification, in which only limited linguistic knowledge, such as the phonotactics of a language, might be enforced [129].

3.1.2 Speech Parameterization

The first stage is to analyze acoustic samples into frames (see Equation 3.1): to actually produce the vector sequence $(o_1, ..., o_t)$ for further analysis. A diagram of how to construct a signal processing front end using currently available techniques is illustrated by the system diagram of Figure 3.3. Here, signal processing can be further broken down into stages:

- First, the raw waveform is preprocessed - at this stage by any combination of a Wiener filter (to remove noise) [130, 131, 132] and preemphasis

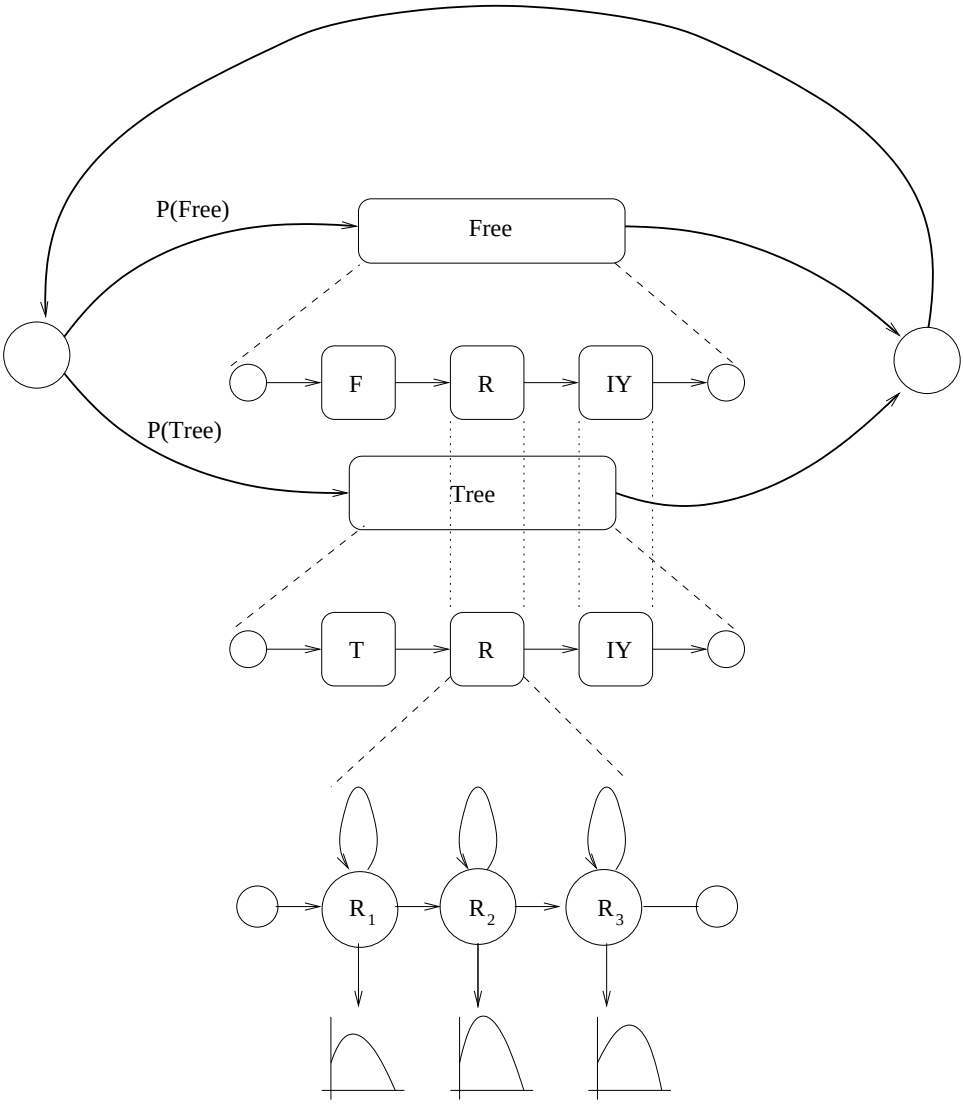


Figure 3.2: Example of an embedded word graph. In this example, we have monophone models shared across a two-word unigram loop grammar.

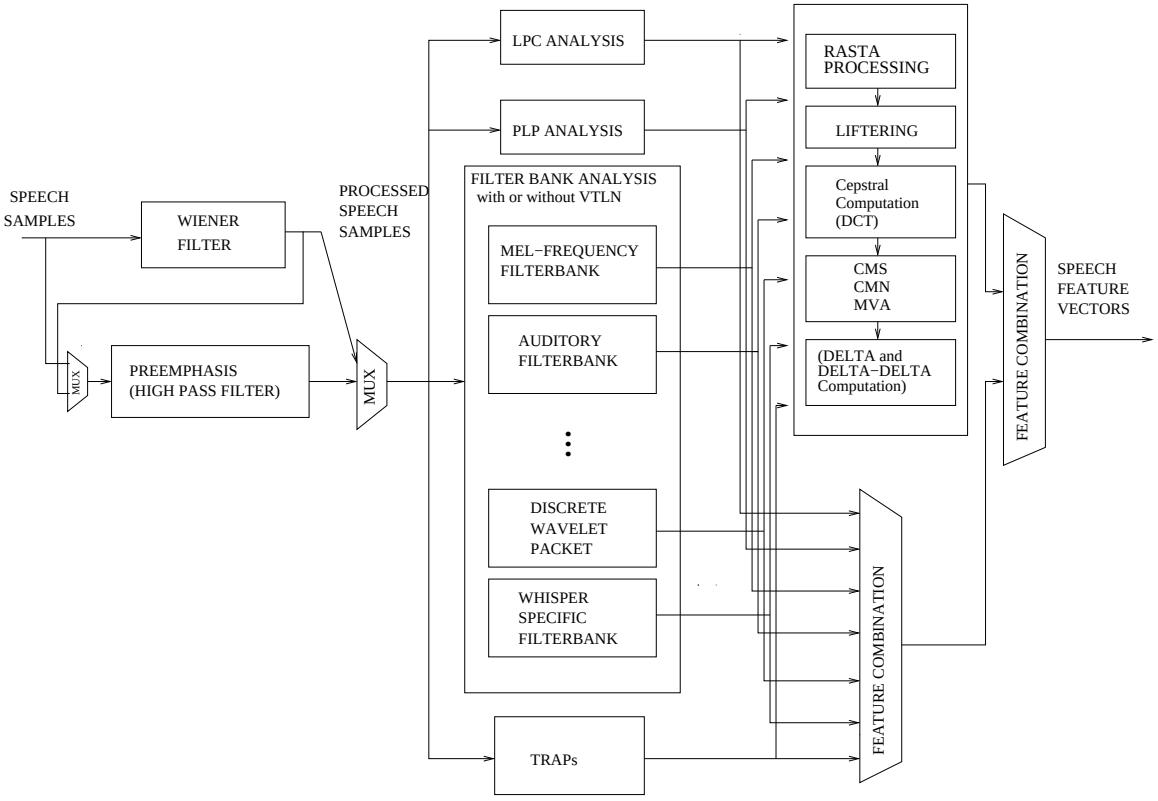


Figure 3.3: A signal processing front end for speech recognition.

(to negate the zero caused by the glottis and vocal tract). The result is a "cleaned-up" version of the acoustic signal, on which recognition should now be easier to perform.

- Next, the processed samples can be fed to some sort of frame-based analysis; this is usually some kind of filter bank — in its simplest form, a Hamming window coupled with the short-time Fourier transform [8] suffices, even though more elaborate techniques such as the Wigner transform [50] could be used. This produces a frame of speech at regular intervals, say every 2 to 10 ms.

- Finally, additional post-processing on the speech vectors can be optionally performed: in particular, speech normalization techniques such as RASTA [133] can be performed, as well as cepstral computation to decorrelate the vector components, so that Gaussian models employing a diagonal variance will be suitable for later acoustic modeling. Cepstral mean normalization [134] can ameliorate volume and "chan-

nel effects" due to the different frequency response of the recording device and the acoustics of the recording environment. Calculation of dynamic features [135] (delta and delta-delta coefficients) can be done to compensate for the inadequacy of having a short time window in a speech frame, unable to capture longer, or more temporal acoustic events. The final output is a sequence of vectors, hand-picked from any combination of available filter banks and frame analysis techniques; in some cases additional processing to reduce the dimensionality of the supervector can be performed [136], resulting in a vector sequence that is now ready to be recognized.

The various signal processing front ends presented here were proposed at different times in the study of machine recognition. Beginning with the earliest, we have:

- The LPC and the LPC cepstra, which can directly capture the resonances of the oral cavity, but will not be able to capture pitch well [33].

- The PLP cepstra, which weights the spectral coefficients output from a filter bank with filters spaced at critical bands from each other, in a way that mimics auditory sensitivity at different frequencies, before taking cepstral computation [137].

- The mel-frequency cepstral coefficients employ frequency warping using the mel-scale. The mel-scale is perceptually scaled to compensate for the behavior of the human auditory system in having differing sensitivities to change in frequency at different frequencies [138]. The mel-frequency spectral coefficients can be generated by overlapping a series of triangular filters at regular intervals along the mel-frequency scale, this is essentially implemented by windowing the magnitude STFT.

- Frame-based analysis using a single filter bank is unable to capture long-term modulations in the speech signal; some of this is thought to be important for the recognition of consonants. The absence of dynamics in a frame by frame approach can be crudely addressed by using delta and delta-delta features [4]. The temporal pattern features (TRAPs) suggested by Hermansky [139, 140] compensate for this by considering features that incorporate a long term temporal slice of the

time-frequency resolution of the signal (i.e. a horizontal window of information centered around the region of interest), as opposed to a single frame (i.e. a vertical slice of the spectrogram).

- The term auditory filter bank seems to refer to several variations of the same basic structure. The design of such filter banks is motivated by our understanding of how the inner ear functions: the behavior of the traveling wave causes the greatest displacement at a specific length along the basilar membrane, which in turn has that mechanical energy transduced by a sharply frequency-selective inner hair cell [141]. Auditory filter banks generally have a filter bank in the initial stage spaced at critical bandwidths along the frequency spectrum, followed by half-wave rectification simulating the inner hair cell response, then low pass filtering simulating the slow temporal response of the spiral ganglion cells. The type of filter bank used could be constant-Q or gammatone, either of which mimics the response properties of the basilar membrane.

- The discrete wavelet packet (DWP) is a [121] variation of the discrete wavelet transform (DWT), in which an arbitrary tree structure is used. This allows us to make arbitrary balances and trade-offs between time and frequency resolution at different parts of the spectrum.

3.2 Acoustic Pattern Recognition Using Hidden Markov Models

A thorough categorization of all the approaches for pattern recognition is beyond the scope of this dissertation — the interested reader is instead referred to the excellent textbook by Duda et al. [142]. As before, we will focus on the hidden Markov model. Over the years, the training of HMMs for speech has evolved to the point of almost being an art - building a good acoustic model involves many, sometimes arcane, tricks. A starting point to produce a good baseline recognizer can be found in the documentation [143] for HTK. The parameters of the model have to be carefully chosen in order to minimize the number of parameters that need to be trained, which at the same time reduces the amount of training data required to build a decent acoustic model.

In the hidden Markov model, observations of the data are conditioned upon an unknown hidden state. A diagram for a typical three-state model is shown in Figure 3.4 [143]. Such a model would be used to model the evolution of observations over the duration of a phoneme. At each time step, the hidden state may evolve to a new state or stay the same. The probabilities of state evolution are model by a matrix A, where $A_{i,j}$ represents the conditional probability of the hidden state going from state i to state j between any particular pair of speech frames, given that we are already in state i. Zero entries represent impossible transitions. A common rule used is to impose the restriction that states evolve from "left to right": here, state 3 does not go back to state 1. The motivation for doing this in phoneme modeling, for instance say a plosive, is so that we might see a progression of feature observations as we enter the stop, release, and vowel onset for the plosive. In this case we should not expect to see the phoneme release or stop phases after vowel onset, and so on and so forth.

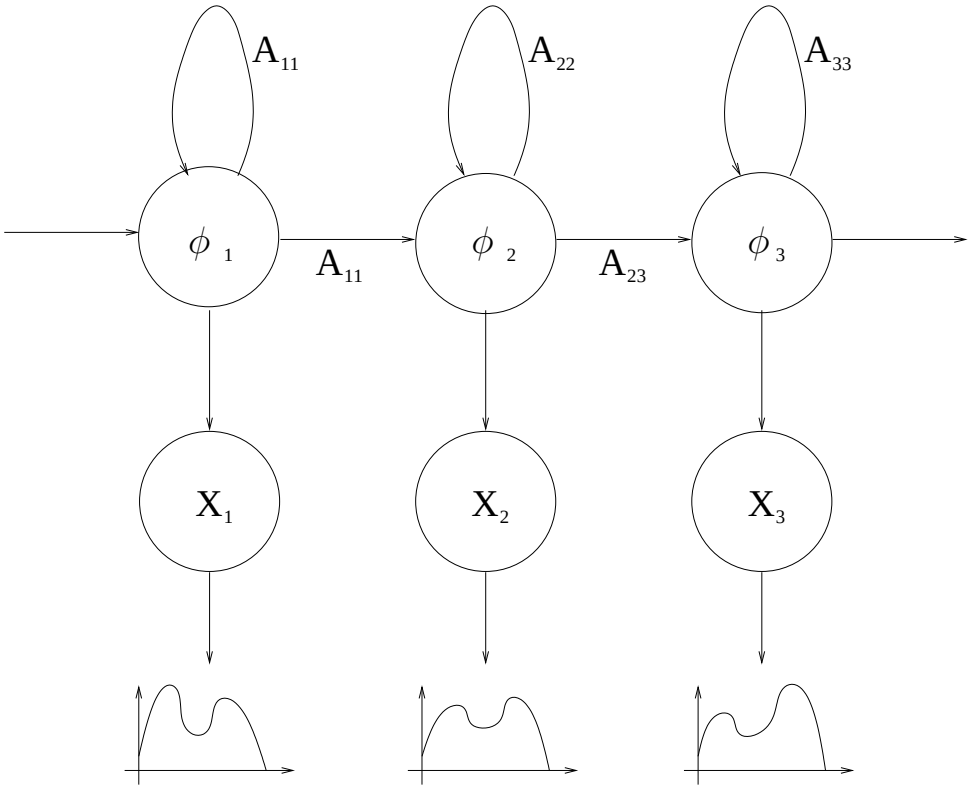


Figure 3.4: Diagram of a three-state left-to-right HMM.

In speech recognition, the observations usually come from a combination of feature extraction techniques mentioned in section 3.1.2, and essentially

give us a stream of feature vectors that change over time. Although it is not necessary, a multivariate mixture Gaussian random variable is often used to model these vector sequences. That is, for the observation x given the specific phoneme s and state i, its likelihood is given by

$$P(x|\lambda_{ac,s,i}) = \sum_{k=1}^{K} w_{s,i,k}(2\pi)^{-\frac{d}{2}}|\Sigma_{s,i,k}^{-\frac{1}{2}}|e^{(x-\mu_{s,i,k})^T\Sigma_{s,i,k}^{-1}(x-\mu_{s,i,k})}, \tag{3.5}$$

which follows a Gaussian mixture distribution with K components, where w's are the weights of each mixture component k, and μ's and Σ's are parameters of Gaussian distributions. Training of the HMM is achieved using the expectation-maximization algorithm [110], which iteratively improves from an initial estimate of the model. The E-step of the algorithm guarantees an increase in the log-likelihood of the data given the model, and thus guarantees convergence.

3.2.1 Triphone Clustering

Coarticulation is often observed in fluent speech: various acoustic parameters of a phoneme may be altered depending on its preceding and succeeding phonemic context [18]. One way to handle this is to explicitly model every likely context for every possible phoneme. These so-called triphones let us assign different probability distributions for the same phone when seen in different contexts [144]. Unfortunately, the use of triphones greatly expands the number of required parameters for the model: a system with originally 39 monophones now has 39^3 or roughly 1500 times the number of parameters to train. This presents a problem in that we would require roughly that many times the amount of data in order to get equally good estimates for the parameters of our multivariate Gaussians. Furthermore, the distribution of triphones is highly non-uniform – there may be some triphones that occur often, but other plausible ones which do not violate phonotactics may not occur at all. In order to alleviate this, parameter sharing through clustering can be applied. The basic principle here is to assume that some triphone states are similar enough in nature so that statistics used for estimating either one of them can be pooled in order to get better statistical estimates.

Clustering may be performed using a top-down approach [145]. We begin

with a set of all triphone states and determine which division most evenly distributes the counts in the data for each state. Divisions happen along lines motivated by linguistic knowledge [146]. For instance, one might divide the states based on a question such as "Is the left context a plosive or not?" The end result of triphone clustering is to produce a set of triphone clusters that allow statistics from particular states of various triphones to be pooled for better estimation. This also reduces the number of trainable parameters substantially.

3.2.2 Word Pronunciation

Hidden Markov models are *generative models*: they belong to a more general class of models known as graphical models; this family includes the HMM's mathematical cousin, the maximum entropy Markov model [125] (MEMM), which directly models observation vectors of speech. Both models are special cases of dynamic Bayesian networks (DBN). Graphical models are powerful in that they can be embedded hierarchically; to make the HMM of a word, one only has to repeatedly concatenate the HMMs for each phone that is contained in the most common pronunciation. Alternate pronunciations can be handled by embedding HMMs corresponding to their representative phoneme sequences in parallel. The decoding of speech can then be performed by considering all possible trajectories through the hidden states, applying likelihoods for each of the observed speech features computed from the raw signal. Usually, the estimation of parameters for HMMs is performed using the Baum-Welch algorithm [147] — a specific incarnation of the expectation-maximization algorithm [111, 148]. Viterbi decoding is usually used to recognize speech [149].

3.2.3 Imposing Language Constraints with n-Grams

In the case of large vocabulary recognition, there is usually little constraint on what can be spoken, although we want to coax the system to try to prefer more meaningful or likely sentences. Under these circumstances, the most pragmatic model to constrain the number of probable utterances is the *n-gram* model [150]. This model works by predicting the next word from the

context of the n-preceding words. The order of the model, n, is the number of previous words that has to be taken into account to make this prediction. Given a sentence S with given a sentence with N words $(w_1, w_2, ...w_{N-1})$, the probability of it occurring is

$$\begin{aligned} P(S) &= P(w_1, w_2, ...w_N) \\ &= \prod_{k=1}^{N} P(w_k|w_k - 1, ...w_1) \\ &\approx \prod_{k=1}^{N} P(w_k|w_k - 1, ...w_{k-n}), \end{aligned} \tag{3.6}$$

where the approximation is due to the assumption that only the next word is influenced by n-preceding models. Some more intricate models improve on this limited context. Trigger models [151] try to model the effect of having pairs of words that may co-occur in a sentence but are far apart, for instance "if" and "then." Bag of words models [152] model long-term contextual information by vectorizing entire sentences or paragraphs and estimating the probability of that vector occurring.

In the case when the actual words in the signal are known, a strict grammar [153] with the words and their alternate pronunciations can be used to decode and automatically find word or phoneme segment boundaries in the speech. Such grammars my be specified in Backus-Naur Form (BNF) and rigidly restrict what can be said.

When the grammar is linear, we end up imposing an exact sequence of words onto the utterance, allowing only variations in pronunciation. This method, known as *forced alignment*, can be used to obtain automatic word or phone segmentation of the speech [154].

3.3 Speaker-Independent Acoustic Modeling

A speaker-independent (SI) acoustic model with the topology described in previous sections can be built from training data obtained from many different speakers. The acoustic model for the set of clustered triphones Φ is the

set of parameters

$$\lambda_{ac} = \{A_{\phi,i,j} : 1 \leq i,j \leq M; w_{\phi,j,k}, \mu_{\phi,j,k}, \Sigma_{\phi,j,k} : \phi \in \Phi, 1 \leq j \leq 3, 1 \leq k \leq K\}. \tag{3.7}$$

This includes parameters such as state transition probabilities $A_{\phi,i,j}$ for going from state i to state j of triphone ϕ, mean $\mu_{\phi,j,k}$ and covariance $\Sigma_{\phi,j,k}$ parameters of Gaussian k in the mixture of state j in triphone ϕ, as well as weights $w_{\phi,j,k}$ for each mixture component. Note that ϕ and j index a clustered triphone state, and might be shared across many different triphones that are deemed similar. Also in our setup we have three-state models for the triphones.

Ideally, there should be enough data from different speakers for each phoneme and context, so that all the parameters are adequately trained. A speaker-dependent (SD) system, however, is trained from data all from the same speaker. The drawback of such a system is that it may not perform well when the speaker is changed. In contrast, a speaker independent system might perform better on new speakers, but since the data for different triphones is drawn from many speakers, they inherently have greater variability, leading to larger covariance values [155]. What this means is that the Gaussians tend to have greater variance, and would give a lower log-likelihood over observed data. In practice it is desirable to obtain a speaker-dependent system for a target speaker when recognizing speech, but the data requirements for adequately training a speaker-dependent system from scratch tends to make this an impossibility. One alternative is to use a small amount of data from the target speaker to modify the speaker-independent system. Such approaches are termed adaptation in the literature.

3.4 Speaker-Adaptation Techniques

There are various speaker adaptation techniques in use, but the most prominent ones are MAP and MLLR [156]. Eigenvoices is an alternative method that developed from a need to perform adaptation using very little data [157]. The three methods differ in both their computation and their ability at adapting acoustic models; MAP requires the most data to function well, but can potentially beat the other two. Eigenvoices can work well with very

little data, but tends to hit a performance asymptote as the amount of data is increased. MLLR on the other hand has properties somewhere between the two.

3.4.1 MAP Adaptation

In the MAP approach [158] we assume a prior distribution for the model λ_{ac}, which is then used to maximize the posteriori probability of observations x, such that

$$\begin{aligned} \lambda_{ac,MAP} &= \underset{\lambda_{ac}}{\operatorname{argmax}}\, p(x|\lambda_{ac})p(\lambda_{ac}) \\ &= \underset{\lambda_{ac}}{\operatorname{argmax}}\, p(\lambda_{ac}|x). \end{aligned} \tag{3.8}$$

In practice, a suitable prior for the model is an initial estimate of the model, which can be obtained through the Baum-Welch algorithm. For speaker adaptation, this is simply the original speaker-independent acoustic model's parameters. The solution to this has to be iteratively estimated using the EM algorithm. Maximization at each iteration ends up as an update, that for most parameters is usually a linear interpolation between the original model and an ML-estimate of parameters obtained from the new data, weighted appropriately by the likelihood of seeing the adaptation data.

This approach only updates mixture components for which data is seen. Thus, a large and varied amount of data that covers parameters of the entire model is necessary in order to get good adaptation results. At the same time, since components are updated individually, new estimates of them are usually very good. This explains the high performance of MAP adaptation when moderate amounts of adaptation data are available.

3.4.2 MLLR Adaptation

Maximum likelihood linear regression (MLLR) can be used to adapt means or variances of the model [159]. MLLR-means works by finding a linear transform on mean parameters, such that the probability of observing the observation data with the new parameters is maximized. Given a set of mean

vectors $\left(\mu_1 ... \mu_k\right)$, we find

$$W^* = \underset{W}{\operatorname{argmax}} P(x|\lambda; W\mu), \tag{3.9}$$

where λ is the model without mean parameters, and W is a linear transform on the mean parameters. In practice, an entire set of W transforms are found, one for each group of parameters belonging to closely related phonemes. The groups can be found by clustering the mean parameters so that phonemes with similar mean cepstral values group together. Similar to MAP, an EM algorithm is required to solve the problem. With this approach, adaptation data seen for one component affects the entire group. This allows entire clusters of phone models to be quickly updated. Thus, less data is required for adaptation. However, this approach does not have the specificity of MAP, and may not perform as well when there is a lot of adaptation data.

It must be noted that when the number of phone clusters is increased the behavior of MLLR steadily approaches that of MAP. Conversely, using a single transform for the entire acoustic model can ameliorate global effects such as channel distortions as well.

MLLR can also be extended to covariance parameters, in which case another linear transform is found that will maximize the likelihood of seen adaptation data. A special case of this is constrained MLLR (CMLLR) [160] , in which both transforms for means and variances are restricted to be the same. In the case of CMLLR, since the exponent for the multivariate Gaussian can be written as $(x - W\mu)^t (W\Sigma)^{-2} (x - W\mu) = (W^{-1}x - \mu)^T (\Sigma)^{-2} (W^{-1}x - \mu)$, we observe that this is effectively computing a projection on the feature vectors. When there is only a single global transform, this is precisely what CMLLR does; when a set of transforms are used, this uses a different linear transform on the feature space depending on the phoneme. Note that if we use the means to actually perform clustering, what this does is produce a mesh-like mapping from one feature space to another.

3.4.3 Eigenvoice Methods

Eigenvoices are a relatively new technique that emerged as a response to the need for "fast adaptation" [161, 157]. The basic idea is to model the variability between different speakers as a vector subspace. First, the mean

vectors $\mu^{(i)}_{\phi,j,k}$ of speaker i's speaker-adapted model are concatenated in a particular order into a supervector $\mu^{(i)}$. These supervectors collectively span a speaker subspace and can be analyzed with principal components analysis (PCA). The matrix

$$U = \left(\tilde{\mu}^{(1)} \quad ... \quad \tilde{\mu}^{(i)} \quad ... \quad \tilde{\mu}^{(K)},\right) = E\Omega V \tag{3.10}$$

gives us a set of eigenvoices $E = (e_0, e_1, ...e_L)$, $L < K$ that characterize the speaker space. Usually, the first eigenvoice e_0 would correspond closely with the speaker-independent mean vector. The means of any speaker-dependent model can now be expressed as

$$\mu^{(i)} = e_0 + \sum_{l=1}^{L} w_l^{(i)} e_l, \tag{3.11}$$

that is, a weighted sum of these eigenvoices. In practice, the matrix U is very large, and performing a full singular value decomposition is very expensive; it is better to compute just the eigenvoices for the M-dimensional subspace that we want. One approach to do this is to use the probabilistic principal components analysis (PPCA) [162], which does exactly that using EM.

In order to adapt to a new speaker, we simply compute new weights — contributions in each eigenvoice — from the adaptation data, and modify the speaker independent model accordingly. Estimation of the new weights is again accomplished using an EM based algorithm, in this case the maximum likelihood eigen-decomposition (MLED). The eigenvoice method modifies the entire parameter set at once, and thus requires very little data to do adaptation. However, the coarse granularity of the updates means that it quickly hits a limit on performance even as more data becomes available.

3.4.4 Combining Methods

In practice, several iterations of EM are required for the various algorithms before the results converge and we get good recognition results. One approach is to combine different approaches to leverage the benefits of each method [163]. This generally works by using a technique that updates at a coarser granularity, followed by a finer technique. For instance, using eigen-

voices followed by MAP, or MLLR followed by MAP, can improve results dramatically over just one method alone.

3.5 Summary

This chapter has outlined the basics behind speech recognition technology. We gave a general framework for the most well studied and advanced methods in use today. We have also specified the topology of the recognizer and acoustic model that we used in this thesis. The three approaches to adaptation will be further explored in Chapter 6 when we consider the problem of adapting normal speech acoustic models to whispered speech acoustic models.

CHAPTER 4

WHISPERED SPEECH CORPORA

Research in whispered speech has been hampered by the lack of large publicly available corpora. A carefully organized and systematically constructed corpus is not only a valuable resource, but a necessary precursor to any meaningful work.

Two different corpora were collected to support different objectives in research. The Whispered TIMIT corpus (wTIMIT) is designed to satisfy the unique needs of constructing large vocabulary speech recognizers, and thus is styled after popular large speech corpora used for this purpose. The corpus is designed to be phonetically balanced, and sufficiently large to support the statistics needed for training acoustic models in speech recognition.

The second corpus, the Whispered Modified Rhyme Test, is designed to help us understand some limits in whispered communication, and thus resembles corpora used in intelligibility tests. The Modified Rhyme Test was specifically chosen as it is a widely recognized test of speech channel intelligibility [164].

4.1 The Whispered TIMIT Corpus

The Whispered TIMIT corpus is modeled after the TIMIT corpus [165], commonly used to study automatic recognition of phonemes. The wTIMIT is a systematically organized collection of paired whispered and spoken utterances produced by several speakers. Collection proceeded in two phases - the first phase consisted of utterances from 20 Singaporean speakers, the second phase consisted of utterances from 28 North American speakers. This resulted in two subsets that differ only in accent. Detailed information on the speakers of each subset is shown in Tables 4.1 and 4.2.

All recordings were made in an audiometric booth using an MX-2001 direc-

Table 4.1: Speakers in first phase collection (collected at NUS).

subject id	gender	age	languages	subject id	gender	age	languages
001	F	25-30	EN,MA	011	M	20-25	EN,MA
002	F	20-25	EN,MA	012	M	20-25	EN,MA
003	M	25-30	EN,MA	013	F	20-25	EN,MA
004	M	25-30	EN,MA,GE	014	M	20-25	EN,MA
005	M	20-25	EN,MA	015	M	15-20	EN,MA,JP
006	F	20-25	EN,MA	016	M	20-25	EN,MA
007	M	15-20	EN,MA	017	M	25-30	EN,MY
008	F	20-25	EN,MA	018	F	20-25	EN,MA
009	F	20-25	EN,MA	019	F	25-30	EN,MA
010	M	25-30	EN,MA,GE	020	M	25-30	EN,MA,JP

KEY: EN - English; MA - Mandarin; JP - Japanese; MY - Malayalam; GE - German

Table 4.2: Speakers in second phase collection (collected at UIUC).

subject id	gender	age	languages	subject id	gender	age	languages
101	F	15-20	EN,SP	102	F	20-25	EN,TH
103	M	20-25	EN	104	F	15-20	EN,GE,YB
105	F	20-25	EN	106	M	20-25	EN,SP
107	M	15-20	EN	108	F	15-20	EN
109	F	15-20	EN,GJ	111	M	30-35	EN
112	F	15-20	EN,FR	115	M	30-35	EN
116	F	15-20	EN,FR,PO	117	M	20-25	EN
118	M	20-25	EN	119	M	15-20	EN
120	F	20-25	EN	121	M	20-25	EN
122	M	25-30	EN,JP	123	F	15-20	EN,GE
124	M	20-25	EN,SP	125	F	15-20	EN,GE
126	F	20-25	EN	127	F	30-35	EN
128	M	15-20	EN	129	F	35-40	EN
130	F	20-25	EN,SP,FR	131	M	15-20	EN

KEY: EN - English; SP - Spanish; PO - Polish; TH - Thai; FR - French; GE - German; JP - Japanese; YB - Yoruba

tional condenser microphone. The microphone was adjusted to be 6 inches away from the speaker's mouth, and tilted a little away to avoid puffs of air hitting the microphone. During whispering, the speaker was told to move closer to the microphone in order to obtain a better dynamic range for recording. Each speaker was requested to both whisper and read a set of 450 prompts. These prompts were obtained from the phonetically balanced section of the TIMIT corpus, and thus cover the most likely phonetic contexts encountered in spoken English. Prompts were alternately read and whispered in sets of 50 so as to avoid speaker fatigue. As far as possible, poorly articulated sentences, mispronounced words and disfluent utterances were rejected and re-recorded, but a minute number of such sentences still made it through quality-control.

4.2 The Whispered Modified Rhyme Test Corpus

The Modified Rhyme Test is an intelligibility test designed to quantify speech communication over spoken channels. This is done by conducting a six-way identification test over a 50 sets of words. The words are all monosyllabic; 25 sets of them differ only in the word-initial consonant, and the other 25 differ only in the word-final position. These word-sets are tabulated in Tables 4.3 and 4.4. Note that there are only 273 distinct words in the set because some words are shared between sets.

Our corpus consists of each word of the set embedded in the carrier sentence "Can you say WORD now." The data was collected in an anechoic chamber with the same MX-2001 directional condenser microphone. For data collection we used a protocol similar to that used with the wTIMIT data set. In all, 28 native North American speakers were recorded this way. A total of 15,180 utterances were left after removing noisy and disfluent utterances.

4.3 Acoustic Analysis

Many differences between normal and whispered speech are well documented in the literature. Most prominently, whisper is commonly claimed to have

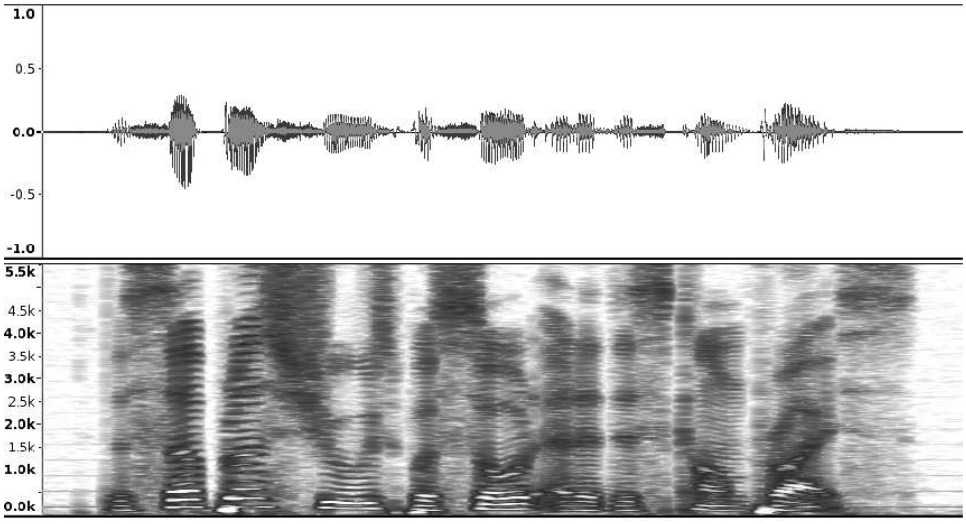
- reduced spectral tilt,

Table 4.3: List of word-initial question sets in the Modified Rhyme Test.

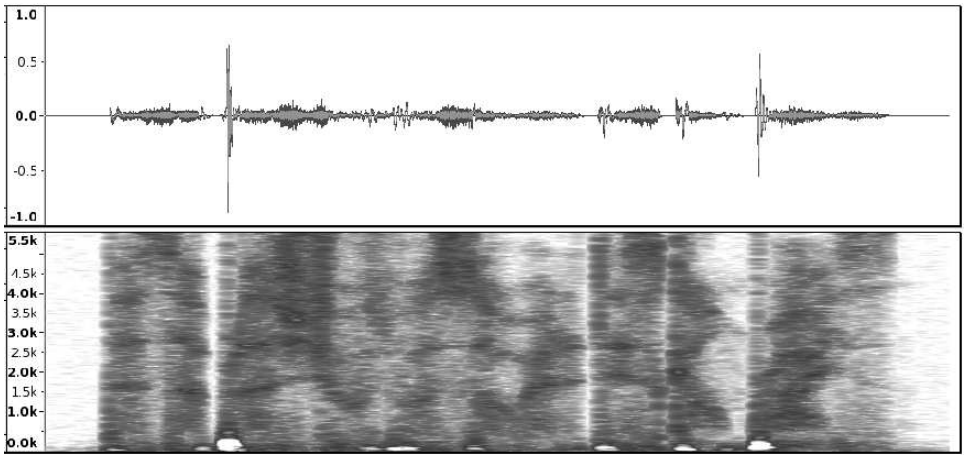
bale	gale	male	pale	sale	tale
bang	fang	gang	hang	rang	sang
bark	dark	hark	lark	mark	park
beat	feat	heat	meat	neat	seat
bed	fed	led	red	shed	wed
bent	dent	rent	sent	tent	went
best	nest	rest	test	vest	west
big	dig	fig	pig	rig	wig
bill	fill	hill	kill	till	will
bit	fit	hit	kit	sit	wit
boil	coil	foil	oil	soil	toil
book	cook	hook	look	shook	took
bun	fun	gun	nun	run	sun
bust	dust	gust	just	must	rust
came	fame	game	name	same	tame
cold	fold	gold	hold	sold	told
cop	hop	mop	pop	shop	top
day	gay	may	pay	say	way
den	hen	men	pen	ten	then
din	fin	pin	sin	tin	win
dip	hip	lip	rip	sip	tip
eel	feel	heel	keel	peel	reel
got	hot	lot	not	pot	tot
jaw	law	paw	raw	saw	thaw
kick	lick	pick	sick	tick	wick

Table 4.4: List of word-final question sets in the Modified Rhyme Test.

bad	bath	back	ban	bass	bat
cud	cud	cuff	cup	cuss	cut
did	dig	dill	dim	din	dip
dub	dud	dung	dug	duck	dun
fib	fig	fill	fin	fit	fizz
kid	king	kick	kill	kin	kit
mad	math	man	map	mass	mat
pad	path	pack	pan	pass	pat
pig	pick	pill	pin	pip	pit
pub	puff	puck	pun	pup	pus
sad	sag	sack	sap	sass	sat
sub	sud	sung	sum	sun	sup
tab	tang	tack	tam	tan	tap
bead	beach	beak	beam	bean	beat
buff	bug	buck	bun	bus	but
cake	came	cane	cape	case	cave
lace	lake	lame	lane	late	lay
pace	page	pale	pane	pave	pay
race	rake	rate	rave	raze	ray
safe	sake	sale	same	sane	save
seed	seethe	seek	seem	seen	seep
sing	sick	sill	sin	sip	sit
heave	heath	heal	heap	hear	heat
peace	peach	peak	peal	peas	peat
tease	teach	teak	teal	team	tear



(a) Normal speech



(b) Whispered speech

Figure 4.1: Spectrograms of normal and whispered speech.

- longer syllables,
- altered (usually raised) formant positions.

In this chapter, we describe some measurements made of the acoustic signal to support these claims.

4.3.1 Waveform and Spectrogram Differences

Figure 4.1 shows the waveform and wide-band spectrogram of normally spoken and whispered versions of the same utterance: "The surplus shoes were

sold at a discount price." Whispered speech can be characterized temporally by the sudden bursts in the signal and noise-like transients. From the spectrogram, we observe that formants are less spectrally peaked and striations typically associated with glottal vibration are absent.

4.3.2 Reduced Spectral Tilt

The spectral quality of speech can be characterized by its long-term average spectrum (LTAS) - this is computed by averaging the magnitude squared of spectral bins obtained from a regularly windowed discrete Fourier transform of the signal. In other words, the spectral bins are

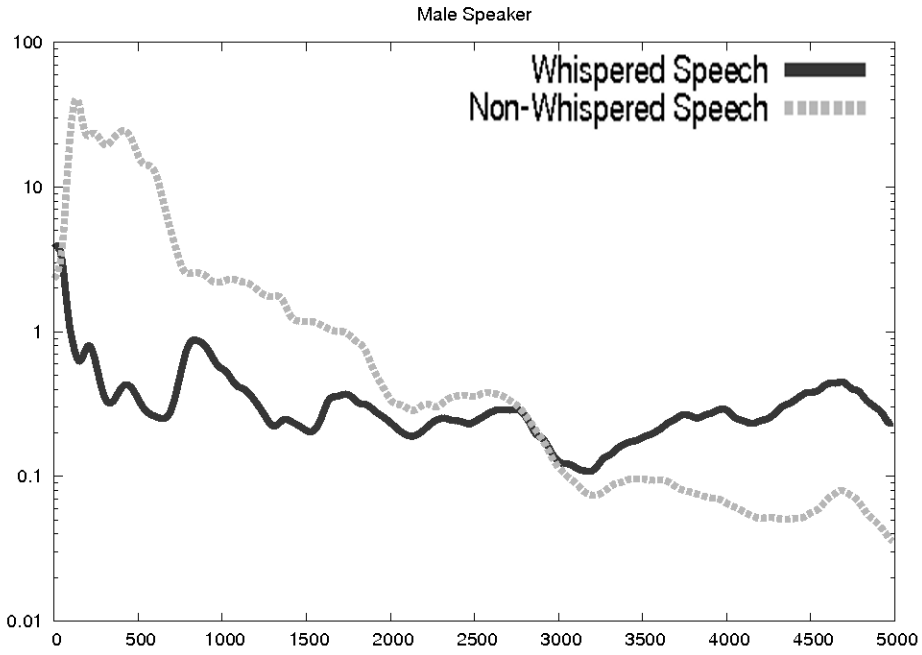
$$\begin{aligned} H_k &= \frac{1}{M}\sum_{m=1}^{M} |DFT_k(x_{n+mT}w_n))|^2 \\ &= \frac{1}{M}\sum_{m=1}^{M} |\sum_{n=1}^{N} x_{n+mT}w_n e^{\frac{\pi*n*k}{N}}|^2, \end{aligned} \tag{4.1}$$

where M is the number of overlapping windows of the signal, T is the window step, and w_n is a suitable windowing function of size N. For our computation, T was chosen to be $\frac{N}{4}$, N is 4096 to be sufficiently wideband. At a sampling rate of 16 kHz, this gives a frequency resolution of around 2 Hz per bin.

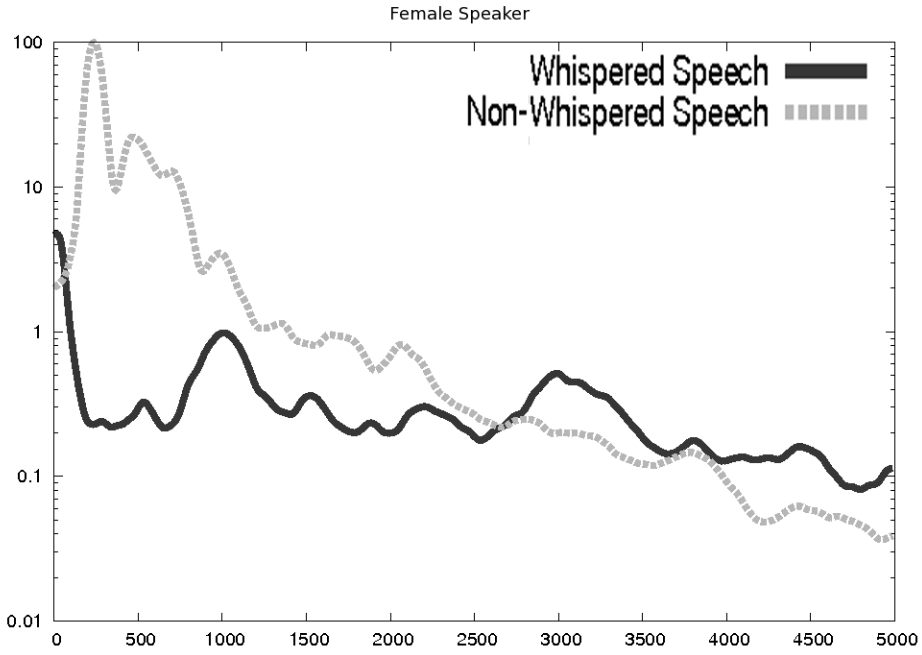
The LTAS of whispered and normal speech from a single male and a single female speaker are contrasted in Figure 4.2. There is little difference in their general shape between the spectra for different genders. Normal speech has a much stronger energy in the low frequency bands as opposed to high frequency bands - it has a greater spectral tilt. This is readily explained by the glottal excitation in normal speech being much more energetic in the low pass regions.

4.3.3 Formant Shift in Vowels

We took measurements of the first three formants of three vowels /a,i,u/ found in the same context from the wTIMIT corpus. Two utterances containing these vowels were drawn from every speaker in the corpus. The first, "A huge power ([p *a* w]) outage rarely ([r ær *i*])," provided the /a,i/ vowels, and the second utterance, "Does Hindu [h i n d *u*] ideology worship cows,"



(a) Male speaker



(b) Female speaker

Figure 4.2: Log-spectral plots for whispered and normal speech. Lighter line corresponds to whisper, darker line corresponds to normal speech.

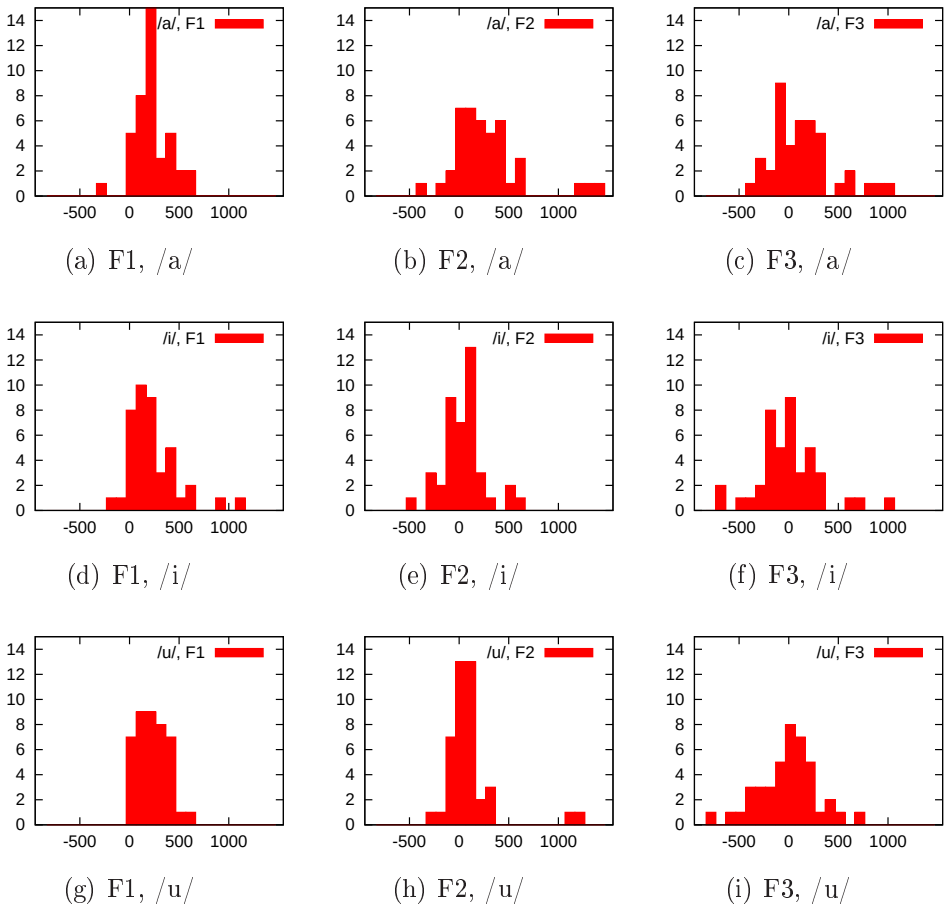
provided /u/. Formant measurements were aided with help of the computer program `Praat` [166]. Temporal locations were found by visual inspection of the spectrogram followed by audio verification. In cases where the algorithm failed to track formants correctly, manual correction was applied. The formant tracks for whispered speech tend to be broader, but they also tend to be far weaker. Occasionally, formant tracks fall below a nominal amplitude and disappear. Some estimation was applied to get reasonable values, but in cases where no reasonable estimate made sense the speech token was discarded.

Formant frequencies are tabulated in Appendix A. In fluent speech we have coarticulation effects and mild mispronunciations due to lazy articulation; thus we may not necessarily get canonical values. However, cursory examination of the data suggests an upward shift of the first formant with whisper, comcomitant with the apparent shortening of the vocal tract. The second and third formant shifts depend on the vowel used.

We compute the change in frequencies of the formants as we go from normal to whispered speech for each speaker. These values are plotted in the histograms shown in Figure 4.3. We can see that on average F1 shifts upwards by 200 Hz for all three vowels. This upward shift is also observed for /a/, but the behavior of F2 and F3 does not appear to change on average for /i,u/. Our results are consistent with findings from [69].

4.3.4 Longer Syllabic Length

Forced alignment using our trained acoustic models was used to phonetically segment the corpora. The average phone lengths were computed from the resulting transcription. As shown in Figure 4.4, whispered phones tend to be longer. This also means that syllables for whispered speech in our corpus tend to be longer, as is reported by many others in the literature.

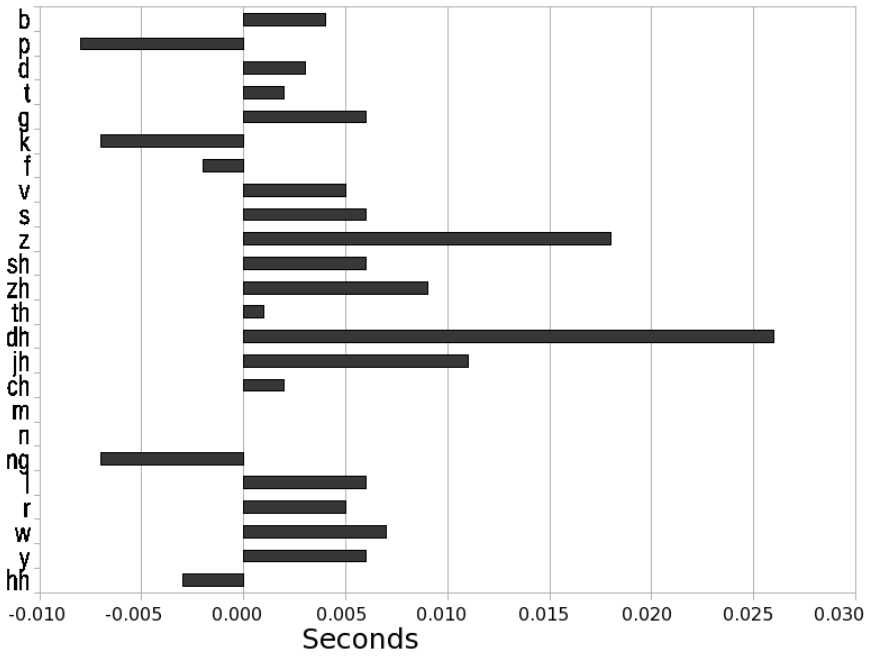


(a) F1, /a/ (b) F2, /a/ (c) F3, /a/

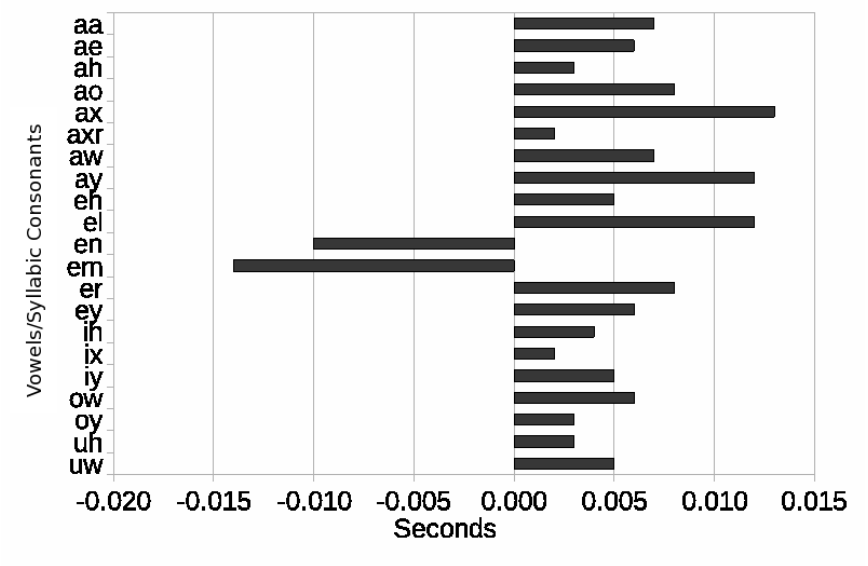
(d) F1, /i/ (e) F2, /i/ (f) F3, /i/

(g) F1, /u/ (h) F2, /u/ (i) F3, /u/

Figure 4.3: Change in formant frequencies going from normal to whispered speech. Horizontal axis is frequency change (Hz), vertical axis is counts.



(a) Consonants



(b) Vowels

Figure 4.4: Average increase in phone length due to whispering.

CHAPTER 5

THE PERCEPTION OF WHISPERED SPEECH

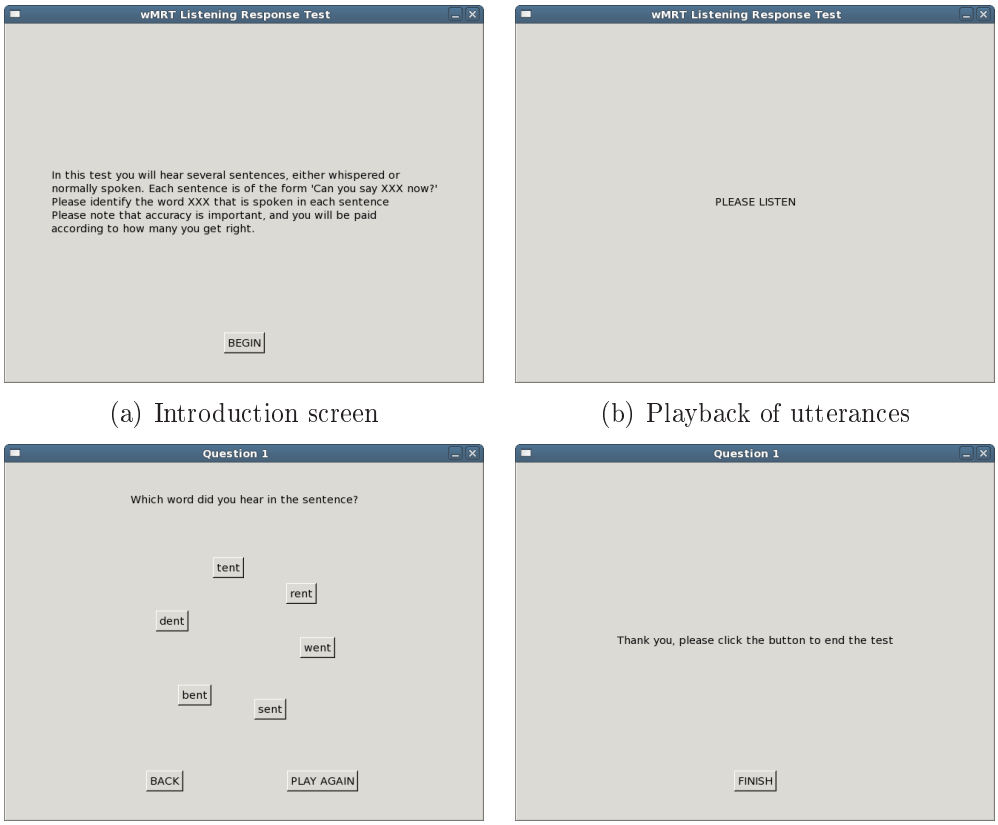
How well does whispering work as a means of communication? Tartter [76] provides confusion matrices of "nonsense" whispered CV syllables, which give us some indication of the answer. Our first experiment complements Tartter's by measuring accuracy of correct identification at word-level contexts.

5.1 Experiment Design of the Whispered Modified Rhyme Test

The test material was drawn from 27 speakers of the wMRT corpus. From this, 20 test sessions, each consisting of 600 questions, were constructed. Each question corresponded to one of the 50 possible word sets of wMRT, with one of the six words in the word set as the correct answer, and could either be read or whispered. Each session consists of utterances drawn at random from the entire corpus, but care was taken to ensure that both the number of male and female speakers, and the number of whispered and unwhispered utterances, were balanced. As far as possible the utterances from different sessions were mutually exclusive – almost all utterances were only used once in the entirety of the testing. Utterances were normalized by root mean squared power, and care taken to ensure that no clipping occurred. The order of presentation of question sets in the session was completely randomized during playback.

A total of 10 male and 10 female listeners, one for each test session, served as subjects in the perceptual experiment. These subjects were paid for their time. Every subject was a native speaker of English from around the Midwest and most were between the ages of 18 to 25; they had at least a high school education. The tests were conducted in an audiometric booth using a custom software program.

Before the start of testing, each subject would be briefed on the test pro-



(a) Introduction screen

(b) Playback of utterances

(c) Presentation of response choices

(d) Ending screen

Figure 5.1: Screenshots from testing program.

cedure and taught how to use the testing software. Subjects were also told to be as accurate as possible, and high accuracy would be rewarded with bonus payment. The software, shown in Figure 5.1, played each utterance in turn and then displayed a selection screen with six buttons – one for each word in the question set. Subjects would listen through a pair of headphones and pick one of the buttons in response. As shown in Figure 5.1, subjects were allowed to replay the utterance as many times as needed to identify the target word. They were also allowed to go back to the previous utterance in case of a misclick, and to adjust the volume of the presentation if it was too loud or too soft.

5.2 Human Perception in the Whispered Modified Rhyme Test

Overall results were 98.9% accuracy on unwhispered speech, and 94.9% on whispered speech. It should be noted that where we do not have 100% accurate identification with phonated speech, any remaining error could be due to any number of factors such as varying accent, varying level of articulateness of speakers, or varying perceptual capabilities of listeners. Other factors such as speaker or listener fatigue may come into play, but this should be ameliorated by the randomized testing as well as the relatively short duration, roughly one hour, of each test and recording session.

Table 5.1: Per listener wMRT identification accuracies.

	Normal Speech		Whispered Speech	
Listener ID	Word-Initial	Word-Final	Word-Initial	Word-Final
001 (M)	100.000	98.000	96.667	95.333
002 (F)	100.000	96.667	95.333	98.000
003 (M)	100.000	94.667	96.000	91.333
004 (F)	100.000	96.000	96.667	94.000
005 (F)	100.000	98.667	97.333	96.000
006 (F)	100.000	97.333	95.333	86.667
007 (F)	100.000	100.000	96.667	98.000
008 (F)	100.000	98.000	94.631	94.667
009 (M)	100.000	99.333	94.667	90.000
010 (F)	98.667	98.000	98.667	91.333
011 (F)	100.000	96.667	93.333	91.333
012 (F)	99.333	98.000	95.333	95.333
013 (M)	100.000	96.667	92.000	91.333
014 (F)	100.000	97.333	98.000	94.667
015 (M)	100.000	100.000	98.000	94.667
016 (M)	100.000	99.329	98.000	94.667
017 (M)	100.000	98.000	96.667	96.000
018 (M)	98.667	98.667	96.000	93.333
019 (M)	100.000	99.333	96.000	94.000
020 (M)	100.000	99.333	97.333	92.667

The effects of such factors are demonstrated clearly when examining identification accuracies per listener, shown in Table 5.1, and accuracies per speaker, shown in Table 5.2. A paired samples t-test using the accuracy values per listener between different speaking styles and different in-word lo-

cation contexts was conducted, and the difference in different speaking style and context was found to be statistically significant. Similarly, the differences in accuracy per speaker were also found to be statistically significant.

Table 5.2: Per speaker wMRT identification accuracies.

	Normal Speech		Whispered Speech	
Speaker ID	Word-Initial	Word-Final	Word-Initial	Word-Final
103 (M)	98.319	92.806	95.000	91.057
105 (F)	100.000	94.545	98.131	85.455
106 (F)	97.059	98.925	92.381	98.780
107 (M)	100.000	96.748	92.105	92.647
108 (M)	100.000	97.581	96.377	90.909
109 (M)	100.000	99.225	98.425	96.639
110 (F)	100.000	100.000	98.246	96.703
111 (F)	100.000	94.792	97.980	88.421
112 (M)	100.000	99.107	92.248	95.935
113 (F)	100.000	100.000	94.231	97.938
114 (F)	100.000	99.083	100.000	94.262
115 (M)	100.000	100.000	95.082	97.458
117 (M)	100.000	100.000	97.080	93.333
118 (M)	100.000	99.200	96.522	95.614
119 (M)	100.000	99.231	95.690	89.583
120 (F)	100.000	100.000	100.000	99.057
121 (F)	100.000	95.652	92.857	86.916
122 (F)	100.000	100.000	92.857	100.000
123 (F)	100.000	100.000	97.778	97.980
124 (F)	100.000	98.095	97.959	97.170
125 (F)	100.000	97.826	94.048	90.909
126 (F)	100.000	97.872	100.000	94.949
127 (M)	100.000	95.455	96.899	96.800
128 (M)	100.000	96.610	96.970	84.259
129 (F)	100.000	94.624	97.938	95.349
130 (M)	100.000	98.601	94.595	92.969
131 (F)	100.000	100.000	95.833	92.308

Overall accuracies are summarized in Table 5.3. Our results indicate that word-final consonants appear to be more confusable than word-initial consonants. This is unlikely to be due to test design itself, as task-entropies for each case were computed and found to be highly similar. The identification task for word-initial differences was computed and found to be 0.432 bits per question; the task entropy of word-final differences was found to be 0.431 bits per question. As expected, whispered speech is less intelligible than

phonated speech, but it is surprisingly not much worse.

Table 5.4 shows some of the most common confusions that occur in the wMRT. Counts of one were discarded. The most common errors in whisper are /b/ → /p/ confusions: there are 14 counts of `bark` → `park`, 13 of `big` → `pig`, as opposed to 5 counts of `pig` → `big` and 3 counts of `pale` → `bale`. Note that there were a total of 20 presentations (since there were 20 listeners) for each type of confusion (one per listener). Clearly, /b/ → /p/ confusions are more common than the reverse. Perhaps in whisper, the plosives resemble the prototypical unvoiced versions more and thus subjects tend to pick the unvoiced version. Yet some identification seems possible, especially for /d/ and /g/ where at least half of the voiced plosives are correctly identified (e.g. only 3 out of 20 for `dip` → `tip`). This supports the idea that some secondary cues are used besides voicing.

The errors were further analyzed and tagged according to the type of confusion associated with them, be it voicing-related, manner-related or place-related. These categories are not mutually exclusive, but multiply tagged errors are uncommon. The counts of each type of error are shown in Figure 5.2. The percentages are computed out of the total number of times the possible error could have occurred. Errors for normal speech tend to be largely place-related, whereas errors in whispered speech tend to be largely voicing related. In the word-final positions, there are a substantial number of manner and place-related errors in whisper, much more than for word-initial position. In other words, the errors in the word-initial position are largely voicing related. This pattern of errors seems to be reversed in normal speech, as there are more voicing related errors in the word-final position. This seems to suggest that in whispered speech different acoustic cues convey information in the word-initial and word-final position.

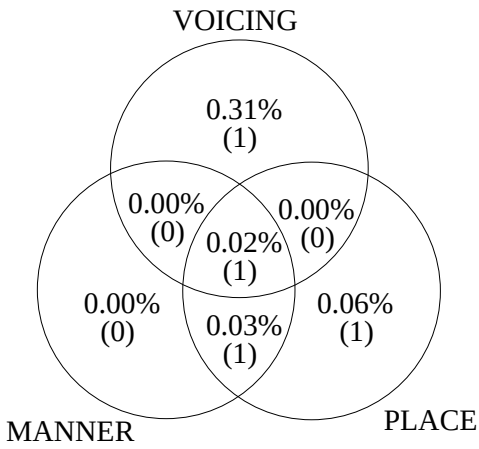
We look at the error confusions associated with the stops in the word-initial and word-final positions. Normalized confusion matrices for whispered

Table 5.3: Human performance for word-initial and word-final consonant recognition.

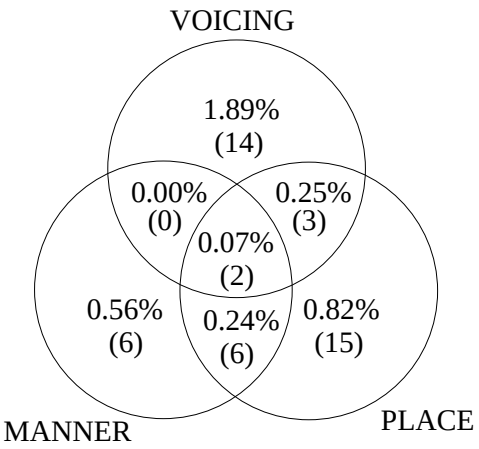
	Speaking Style (%)	
	Normal	Whisper
Word-Initial	99.8	96.1
Word-Final	98.0	93.7

Table 5.4: Most common confusions in human wMRT perception.

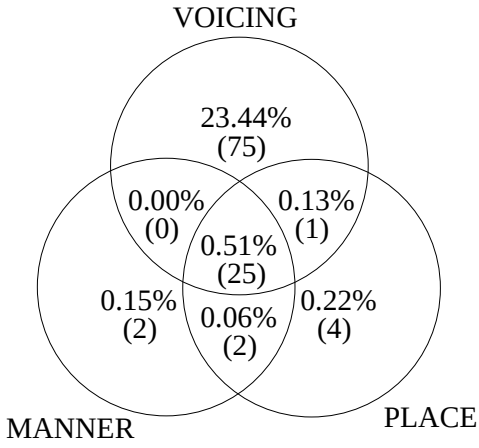
Normal Speech	Whispered Speech			
lane → lay - 3	bark → park - 14	big → pig - 13	bale → pale - 12	gold → cold - 11
heap → heat - 3	din → tin - 8	cud → cut - 8	peas → peace - 6	game → came - 6
cud → cut - 3	den → ten - 6	tab → tap - 5	sub → sup - 5	save → safe - 5
seem → seen - 2	sat → sad - 5	pig → big - 5	bead →beat - 5	bead →bean - 5
sad → sat - 2	tin →din - 4	sup →sub - 4	dug →duck - 4	came →game - 4
pub → pup - 2	seed →seen - 3	sad →sat - 3	raze →race - 3	pay →pane - 3
pane → pay - 2	pat →pad - 3	park →bark - 3	pane →pay - 3	lane →lame - 3
pad → pat - 2	dud →dun - 3	dip →tip - 3	dip →did - 3	dent →tent - 3
dung → dun - 2	bug →buck - 3	vest →best - 2	tip →dip - 2	sud →sun - 2
bat → bad - 2	sing →sin - 2	sin →sit - 2	seethe→seed - 2	seem →seen - 2
ban → bad - 2	same →sane - 2	sag →sack - 2	race →raze - 2	pat →pan - 2
	mat →mad - 2	kit →kid - 2	kid →kit - 2	heave→heath - 2
	heat →heap - 2	heap →heat - 2	fold →hold - 2	fang →bang - 2
	dung →dun - 2	did →din - 2	cut →cud - 2	cold →gold - 2
	cane →came - 2	but →bun - 2	beat →bean - 2	beat →bead - 2
	bath →bad - 2	ban →bad - 2	bad →ban - 2	



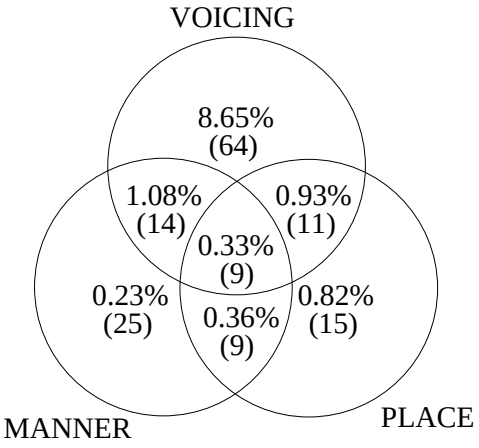
(a) Normal speech, word-initial context



(b) Normal speech, word-final context



(c) Whispered speech, word-initial context



(d) Whispered speech, word-final context

		Voicing-related	Manner-related	Place-related
Normal	Word-Initial	0.026% (2)	0.018% (2)	0.027% (3)
	Word-Final	0.319% (19)	0.184% (14)	0.316% (26)
Whisper	Word-Initial	1.326% (101)	0.259% (29)	2.079% (32)
	Word-Final	1.644% (98)	0.750% (57)	0.534% (44)

Figure 5.2: Categories of errors found in perceptual wMRT.

speech are shown in Table 5.5 for stops: similarly for nasals (Table 5.6), fricatives and affricates (Table 5.7). Since the MRT task involves a six-way forced choice over varying groups of consonants, care has to be taken when normalizing the table. Each diagonal entry of the table corresponds to the probability of correct identification under all contexts. Each off-diagonal entry corresponds to the probability

$$P_{\phi \to \hat{\phi}} = \frac{n_{\phi \to \hat{\phi}}}{N_{\phi \to \hat{\phi}}}, \tag{5.1}$$

where $n_{\phi \to \hat{\phi}}$ is the number of times the stimulus phoneme ϕ gets recognized as $\hat{\phi}$, and $N_{\phi \to \hat{\phi}}$ is the number of times the question set actually allows the particular confusion to be made. Note that with this definition, the row sums will not add to one due to different counts in the denominator for different confusions. Some confusions simply do not occur in the data, as no such word pairs exist. An example of this would be having all presentations of the stimulus "cut" not having "but" as any of the allowed responses; in this case /k/ → /b/, like other such entries, would be marked with an "X."

The corresponding probabilities for normal speech indicate overwhelmingly correct responses and are uninteresting to reproduce here. The "`#`" token is used to represent collectively all other phonemes besides those explicitly named in the table. Nasals, fricatives and affricates are overwhelmingly correct. For plosives in the word-initial position, as well as fricatives and affricates in the word-final position, confusions tend to occur with voiced/unvoiced minimal pairs. There is also a tendency for voiced phonemes to be confused as voiceless, as opposed to voiceless as voiced. All these observations are in agreement with Tartter's [76].

One task of interest is to compute an accuracy of correct voicing transmission. This can be done by considering question sets for which the contrastive phoneme exists as one of the available choices (e.g. presenting the token `put` with `but` as one of the available choices). Accuracy was computed only with data from such question sets, and other data were ignored. Mistakes which did not alter the voicing feature (e.g. /but/ → /gut/) were counted as correct. For the word-initial position, only plosives were involved in such question sets, so our statistics are only valid when considering plosives. The word-final position includes some additional phonemes such as /f/, /v/, /s/ and /z/. Table 5.8 shows the percentage of accurate transmission of voicing

Table 5.5: Perceptual error confusions for stops.

(a) Perceptual, whispered, word-initial confusions.

response / stimuli	/p/	/t/	/k/	/b/	/d/	/g/	#
/p/	0.96	0.00	0.00	0.13	0.00	0.00	0.00
/t/	0.01	0.96	0.00	0.01	0.10	0.00	0.00
/k/	0.00	0.00	0.97	0.00	X	0.15	0.00
/b/	0.65	0.00	0.00	0.86	0.00	0.00	0.00
/d/	0.00	0.25	X	0.00	0.88	0.00	0.00
/g/	0.00	0.00	0.42	0.00	0.03	0.89	0.00
#	0.00	0.00	0.00	0.01	0.00	0.00	0.28

(b) Perceptual, whispered, word-final confusions.

response / stimuli	/p/	/t/	/k/	/b/	/d/	/g/	#
/p/	0.91	0.04	0.01	0.08	0.05	0.00	0.00
/t/	0.02	0.91	0.00	0.00	0.11	0.01	0.01
/k/	0.01	0.02	0.97	0.02	0.00	0.04	0.00
/b/	0.18	0.00	0.00	0.88	0.00	0.00	0.00
/d/	0.01	0.11	0.00	0.00	0.72	0.00	0.04
/g/	0.00	0.01	0.09	0.00	0.00	0.71	0.02
#	0.00	0.01	0.00	0.00	0.02	0.00	0.27

Table 5.6: Perceptual error confusions for nasals.

(a) Perceptual, whispered, word-initial confusions.

response / stimuli	/m/	/n/	/ŋ/	#
/m/	1.00	0.00	X	0.00
/n/	0.00	0.98	X	0.00
/ŋ/	X	X	X	X
/#/	0.00	0.00	X	0.18

(b) Perceptual, whispered, word-final confusions.

response / stimuli	/m/	/n/	/ŋ/	#
/m/	0.96	0.03	0.00	0.00
/n/	0.04	0.95	0.00	0.01
/ŋ/	0.00	0.04	0.96	0.00
#	0.01	0.02	0.00	0.20

Table 5.7: Perceptual error confusions for fricatives and affricates.

(a) Perceptual, whispered, word-initial confusions.

response stimuli	/f/	/θ/	/s/	/ʃ/	/ʧ/	/v/	/ð/	/z/	/ʒ/	/ʤ/	#
/f/	0.97	X	0.00	0.00	X	X	X	X	X	X	0.01
/θ/	X	1.00	0.00	X	X	X	X	X	X	0.00	0.00
/s/	0.00	0.00	1.00	X	X	X	X	X	X	0.00	0.00
/ʃ/	0.00	X	X	1.00	X	X	X	X	X	X	0.00
/ʧ/	X	X	X	X	X	X	X	X	X	X	X
/v/	X	X	X	X	X	0.90	X	X	X	X	0.02
/ð/	X	X	X	X	X	X	0.95	X	X	X	0.01
/z/	X	X	X	X	X	X	X	X	X	X	X
/ʒ/	X	X	X	X	X	X	X	X	X	X	X
/ʤ/	X	0.00	0.00	X	X	X	X	X	X	0.97	0.01
#	0.00	0.00	0.00	0.00	X	0.01	0.00	X	X	0.00	0.21

(b) Perceptual, whispered, word-final confusions.

response stimuli	/f/	/θ/	/s/	/ʃ/	/ʧ/	/v/	/ð/	/z/	/ʒ/	/ʤ/	#
/f/	0.99	X	0.00	X	X	0.05	X	X	X	X	0.00
/θ/	X	0.93	0.00	X	X	0.00	X	X	X	X	0.02
/s/	0.00	0.00	0.97	X	0.00	0.02	X	0.05	X	0.00	0.00
/ʃ/	X	X	X	X	X	X	X	X	X	X	X
/ʧ/	X	X	0.00	X	1.00	X	X	0.00	X	X	0.00
/ v/	0.25	0.10	0.00	X	X	0.91	X	0.05	X	0.00	0.00
/ð/	X	X	X	X	X	X	X	X	X	X	X
/ z/	X	X	0.23	X	0.03	0.00	X	0.88	X	X	0.00
/ʒ/	X	X	X	X	X	X	X	X	X	X	X
/ʤ/	X	X	0.00	X	X	0.00	X	X	X	1.00	0.00
#	0.00	0.00	0.00	X	0.00	0.00	X	0.00	X	0.00	0.21

Table 5.8: Accuracy of voicing feature transmission computed from wMRT result.

	Voicing Transmission (in %)					
Context	/b/	/d/	/g/	/p/	/t/	/k/
Word-Initial	35.0	75.0	57.5	85.0	86.7	90.0
Word-Final	81.7	88.6	91.3	91.7	84.2	95.0
	/f/	/v/	/s/	/z/		
Word-Final	95.0	75.0	95.0	77.5		

per phoneme. The overall accuracy of transmission is 71.528% for plosives in the word-initial position, and 90.526% in the word-final position. Transmission of voicing in the word-final position is 87.482% when the fricatives are included. Overall accuracy of voicing transmission is 79.505%. Our figures are higher than the figure of 64% obtained by Tartter [76], but perhaps this could be explained by the greater amount of information that is conveyed with the full word contexts involved in our recognition task. These results also reaffirm that unvoiced plosives more accurately convey voicing than voiced plosives.

5.2.1 Effect of Gender

In this set of analyses we consider how gender affects identification accuracy. Table 5.9 shows overall identification accuracies for each gender of speaker and listener, for normal and whispered speech in the word-initial and word-final contexts. In order to investigate whether the differences were significant, an independent samples t-test was conducted on accuracy values obtained for individual listeners. The test was conducted for three different pairs of groups, depending on the listener gender, speaker gender and whether there was a gender mismatch.

The analysis found that accuracy differences between different listener gender groups was not significant for both normal and whispered speech in both word-initial and word-final positions. Accuracy differences for when the speaker and listener gender were mismatched and when they were not, were insignificant.

We performed another analysis, this time based on the accuracies computed per speaker. Accuracy differences between different listener and speaker

Table 5.9: Overall identification accuracy for different genders. Numbers in brackets are the total number of questions.

(a) Normal speech, word-initial context

	Listener Gender	
Speaker Gender	Male	Female
Male	99.869 (765)	99.872 (783)
Female	99.864 (735)	99.721 (717)

(b) Normal speech, word-final context

	Listener Gender	
Speaker Gender	Male	Female
Male	98.153 (758)	97.613 (754)
Female	98.516 (741)	97.721 (746)

(c) Whispered speech, word-initial context

	Listener Gender	
Speaker Gender	Male	Female
Male	95.269 (782)	96.016 (728)
Female	97.075 (718)	96.239 (771)

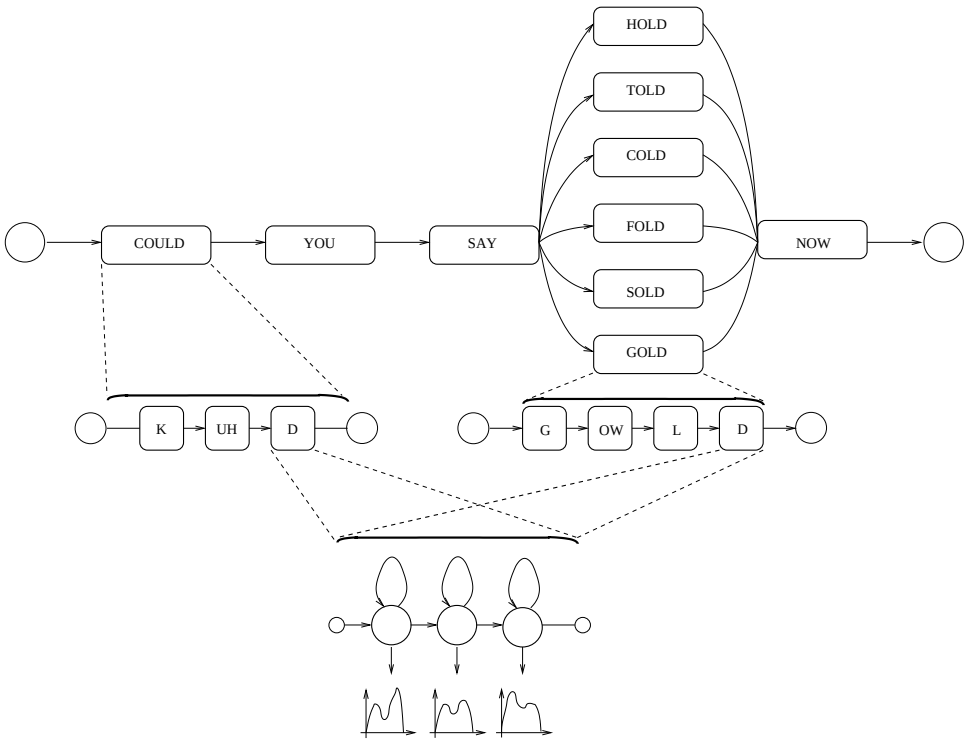
(d) Whispered speech, word-final context

	Listener Gender	
Speaker Gender	Male	Female
Male	92.818 (738)	93.387 (741)
Female	93.832 (762)	94.598 (759)

groups were found to be insignificant. Differences in per speaker accuracy when listener and speaker genders are matched or mismatched are also insignificant. Our results suggest that there is no statistical difference in the articulation of male and female speakers. However, there could be a slight difference in perception between male and female listeners, although their performance is statistically neither better nor worse when listening to normal and whispered speech of the opposite gender.

5.3 Machine Recognition of wMRT Sentences

The wTIMIT trained acoustic models (see next chapter) were used analogously to the perceptual tests in order to determine machine performance on the same task. These acoustic models were context-dependent clustered triphone models. This approach meant that we had an acoustic model inde-




```
$WORD=hold|cold|told|fold|sold|gold;
(<SIL> could you say $WORD now <SIL>)
```

Figure 5.3: Example of a regular grammar for a wMRT question set.

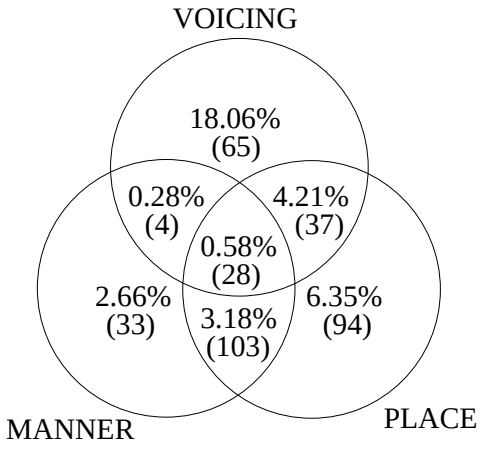
pendently trained from the test sentences in wMRT. Speech recognition was performed using an appropriately constructed regular grammar that contained the carrier sentence and permitted the appropriate word choices. An example grammar is shown in Figure 5.3.

Exactly the same test sets as those presented to human listeners were used. Accuracy values shown in Table 5.10 do not suggest much worse performance in whispered speech compared to non-whispered speech for either machine or human recognition. Machine performance is much worse than human, though more could be done to improve the speech recognition, e.g. by performing speaker adaptation.

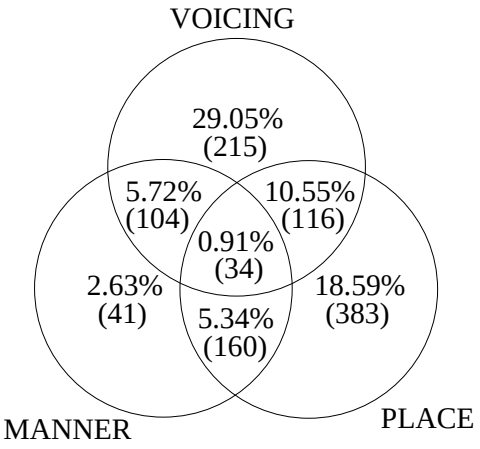
We can perform the same analysis on the errors as we did for the perceptual tests. The errors made by the ASR system are categorized and shown in Figure 5.4. Non-voicing-related errors now occur far more often, and place-related errors seem to occur very often. There appears to be a consistent pattern of errors that occur with both normal speech and whispered speech

Table 5.10: Machine and human performance for whispered and unwhispered speech recognition.

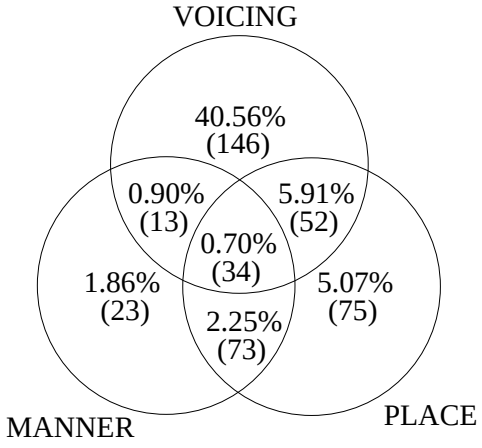
	Speaking Style (%)	
	Normal	Whisper
Human	98.8	94.8
ASR	80.9	77.5



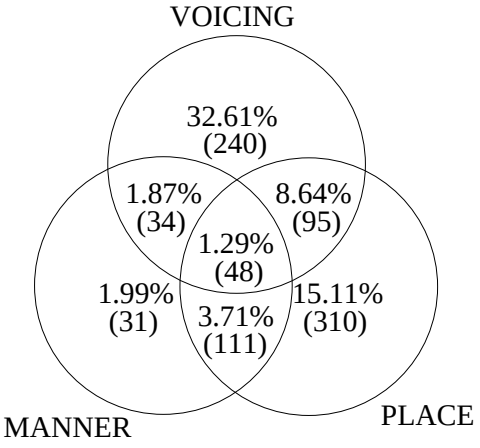
(a) Normal speech, word-initial



(b) Normal speech, word-final



(c) Whispered speech, word-initial



(d) Whispered speech, word-final

		Voicing-related	Manner-related	Place-related
Normal	Word-Initial	1.782% (134)	1.561% (168)	2.510% (262)
	Word-Final	6.346% (469)	3.355% (339)	7.009% (693)
Whisper	Word-Initial	3.260% (245)	1.329% (143)	2.242% (234)
	Word-Final	5.642% (417)	2.216% (224)	5.709% (564)

Figure 5.4: Categories of errors found in ASR wMRT.

– the system does not appear to make mistakes differently for either type of speech.

Tables 5.11, 5.12 and 5.13 show the confusion patterns for the various manner categories. As can be seen, there is much less structure compared with the perceptual results. Notably, the bias in confusions preferring a [voiced] → [voiceless] type of error is absent. However, errors still appear to be chiefly with respect to the voicing distinction. Finally, with the word-final stops, errors occur along one error dimension, either voicing or place, and combination type errors appear to be rarer.

An analysis of voicing transmission now completes our comparison with the perceptual results. These figures are tabulated in Table 5.14. Voicing is transmitted with of 54.931% for word-initial plosives, and 63.330% for word-final plosives. Overall voicing transmission in the word-final position is 64.258%, and overall accuracy of voicing transmission is 59.094%.

The pattern of ASR errors is substantially different from perceptual errors. Surprisingly, some plosives have better voicing transmission in ASR than in human perception. Thus, although ASR accuracy is lower than human accuracy on average, it is not true that humans outperform ASR for every phoneme.

5.4 The Effect of Context in Communication

The perceptual and ASR results give a good idea of how whispered speech carries across information at word-level contexts. Our result is comparable to Tartter's [76]. In the perceptual results, a similar bias to mistake [voiced] → [voiceless] was observed. Machine recognition makes different mistakes from humans, and overall is worse. Interestingly, limiting the listener responses to valid word choices seems to help machine recognition reach the level seen with nonsense CVs. This seems to have a greater effect in human perception – limited task entropy helps recognition even more. Tarttar's result gives us an accuracy of 64% for nonsense CVs; this figure goes up to 94% when word contexts are considered.

Given this result it is uncertain whether or not it is necessary for specific distinctive features to be completely correctly conveyed. The question of how much a distinctive feature contributes to discriminating words is addressed

Table 5.11: Error confusions for stops in ASR wMRT.

(a) ASR, whispered, word-initial.

response stimuli	/p/	/t/	/k/	/b/	/d/	/g/	#
/p/	0.85	0.01	0.00	0.47	0.01	0.00	0.01
/t/	0.05	0.69	0.18	0.09	0.35	0.12	0.01
/k/	0.02	0.07	0.69	0.03	X	0.80	0.02
/b/	0.38	0.01	0.01	0.91	0.00	0.00	0.00
/d/	0.14	0.41	X	0.06	0.62	0.17	0.00
/g/	0.00	0.15	0.05	0.01	0.07	0.89	0.00
#	0.02	0.01	0.00	0.02	0.02	0.02	0.27

(b) ASR, whispered, word-final.

response stimuli	/p/	/t/	/k/	/b/	/d/	/g/	#
/p/	0.69	0.19	0.14	0.43	0.02	0.00	0.01
/t/	0.03	0.76	0.18	0.00	0.06	0.10	0.02
/k/	0.07	0.34	0.72	0.00	0.06	0.01	0.01
/b/	0.30	0.65	0.08	0.43	0.00	0.23	0.03
/d/	0.03	0.70	0.11	0.33	0.22	0.14	0.06
/g/	0.00	0.23	0.50	0.00	0.08	0.26	0.10
#	0.03	0.07	0.01	0.01	0.04	0.02	0.25

Table 5.12: Error confusions for nasals in ASR wMRT.

(a) ASR word-initial confusions.

response stimuli	/m/	/n/	/ŋ/	#
/m/	0.93	0.25	X	0.01
/n/	0.15	0.88	X	0.02
/ŋ/	X	X	X	X
#	0.01	0.01	X	0.18

(b) ASR word-final confusions.

response stimuli	/m/	/n/	/ŋ/	#
/m/	0.69	0.28	0.00	0.02
/n/	0.34	0.73	0.06	0.03
/ŋ/	0.00	0.14	0.85	0.00
#	0.02	0.04	0.00	0.20

Table 5.13: Confusions for fricatives and affricates in ASR wMRT.

(a) Whispered, word-initial.

response / stimuli	/f/	/θ/	/s/	/ʃ/	/ʧ/	/v/	/ð/	/z/	/ʒ/	/ʤ/	#
/f/	0.94	X	0.03	0.00	X	X	X	X	X	X	0.01
/θ/	X	0.80	0.05	X	X	X	X	X	X	0.00	0.05
/s/	0.02	0.00	0.98	X	X	X	X	X	X	0.00	0.00
/ʃ/	0.00	X	X	1.00	X	X	X	X	X	X	0.00
/ʧ/	X	X	X	X	X	X	X	X	X	X	X
/v/	X	X	X	X	X	0.65	X	X	X	X	0.07
/ð/	X	X	X	X	X	X	0.30	X	X	X	0.14
/z/	X	X	X	X	X	X	X	X	X	X	X
/ʒ/	X	X	X	X	X	X	X	X	X	X	X
/ʤ/	X	0.00	0.00	X	X	X	X	X	X	0.93	0.02
#	0.01	0.07	0.00	0.00	X	0.00	0.00	X	X	0.00	0.21

(b) Whispered, word-final.

response / stimuli	/f/	/θ/	/s/	/ʃ/	/ʧ/	/v/	/ð/	/z/	/ʒ/	/ʤ/	#
/f/	0.88	X	0.02	X	X	0.20	X	X	X	X	0.02
/θ/	X	0.46	0.00	X	X	0.80	X	X	X	X	0.09
/s/	0.02	0.00	0.99	X	0.00	0.00	X	0.00	X	0.00	0.00
/ʃ/	X	X	X	X	X	X	X	X	X	X	X
/ʧ/	X	X	0.00	X	0.98	X	X	0.00	X	X	0.00
/v/	0.40	0.00	0.20	X	X	0.65	X	0.00	X	0.00	0.04
/ð/	X	X	X	X	X	X	X	X	X	X	X
/z/	X	X	0.78	X	0.07	0.00	X	0.55	X	X	0.01
/ʒ/	X	X	X	X	X	X	X	X	X	X	X
/ʤ/	X	X	0.50	X	X	0.00	X	X	X	0.50	0.00
#	0.02	0.02	0.00	X	0.05	0.03	X	0.00	X	0.07	0.20

Table 5.14: Accuracy of voicing feature transmission computed from ASR wMRT result.

	Voicing Transmission (in %)					
Context	/b/	/d/	/g/	/p/	/t/	/k/
Word-Initial	60.0	38.8	92.5	53.3	65.0	20.0
Word-Final	66.7	21.4	46.1	55.0	92.5	98.3
	/f/	/v/	/s/	/z/		
Word-Final	80.0	60.0	100.0	22.5		

Figure 5.5: Filtering approach to estimate information transmitted.

by the concept of functional load [167]. Early methods to quantify this involve counting the number of minimal pairs having the particular feature opposition, which can be skewed depending on the probability of occurrence of the minimal pairs. Hockett provides an information theoretic approach which is also adopted by Carter [168] and Surendran and Niyogi [169, 170], which we will paraphrase here. The approach works by measuring the entropy difference between text from a filtered and unfiltered language.

5.4.1 Entropy Loss in Filtered Speech

We begin by considering a hypothetical filter, which conflates contrastive phonemes for a particular distinctive feature. For example, the conflation filter might conflate /b/ and /p/. This is modeled in Figure 5.5. Here, X represents a random process which generates a sequence of speech tokens, specifically phonemes. Since this system is deterministic, the conditional entropy $H(Y|X)$ is 0. Applying the identity for mutual information, $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, gives us the amount of information lost by the conflation of /b/ and /p/ as

$$H(X|Y) = H(X) - H(Y) \tag{5.2}$$

bits. In another sense, this quantifies the information conveyed by having the particular distinction. Suppose we sample a sequence of tokens x from the process X. If we model X and Y with an appropriate generative model trained from the corpus x, we can consider the difference between entropies of those models instead to be an estimate of information transmitted in bits

$$H(X|Y) \approx \frac{\log_2 p_X(x) - \log_2 p_Y(y)}{n}, \tag{5.3}$$

where $p_X(x)$ is the probability of token sequence x under the model X, and n is the number of tokens in x. The functional load of the distinctive feature is given by

$$\begin{aligned} FL(feature) &= \frac{H(X) - H(Y)}{H(X)} \\ &= \frac{H(X|Y)}{H(X)} \times 100\%, \end{aligned} \tag{5.4}$$

as a percentage of the bits per token that are actually involved in transmitting the feature conflated out. This gives us a procedure to estimate the amount of information transmitted by the contrastive pair, using the respective n-gram models trained from those pieces of data.

We processed the `APW` segment of the English Gigaword corpus, and obtained a unigram count of words. This word list was then used with the Sequitur grapheme to phoneme algorithm [171] in order to produce a dictionary of phonemic spellings. The algorithm was trained using the public phonetic lexicon, `cmudict`, from CMU. Resubstituting all words in the corpus with their phonemic spellings gave us a corpus of phonemic tokens to work with. We now apply our procedure to the phonetized corpus, using different conflation filters. First, all vowels were conflated to a single token. The SRI language modeling tools were then used with the original corpus to compute an n-gram language model, from which the corpus cross-entropy is computed. The conflation filter produces a distorted text, from which a similar model is trained and a cross-entropy computed. The complexity of the language model used for computation, that is, the n of the n-gram, relates to the length of the context we are studying. The number of bits per token needed to transmit the contrastive feature can be computed from the cross-entropy of the undistorted corpus computed with the undistorted model. Dividing this by the cross-entropy of the undistorted corpus allows us to find out what percentage of the information is carried by voicing.

Results for unigrams, bigrams and trigrams are summarized in Table 5.15. The numbers in parentheses are functional loads. We observe that our unigram cross-entropy for the baseline is similar to Shannon's value of 2.6 for English [172]. Information transmitted by being able to distinguish various phoneme pairs varies from phoneme to phoneme, with /d/ and /t/ being the least informative among plosives, and /s/ and /z/ most informative among

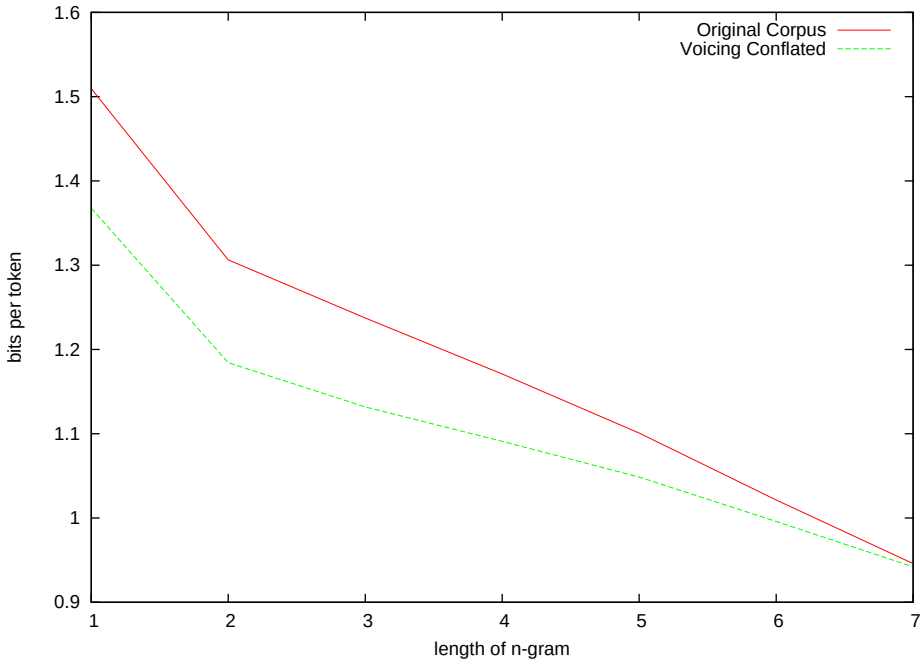


Figure 5.6: Reducing entropy of the language with increased context.

fricatives and affricates. As the model complexity is increased, the amount of information conveyed is reduced; i.e. context effects reduce the need for voicing distinctions to convey information. This is also illustrated by Figure 5.6, which shows how the number of bits required to transmit the language is reduced as context length is increased. With context of 7 phone tokens, voicing only conveys 0.407% of the information in the language. The results seem to suggest that context may obviate all but 1% to 5% of the information carried by voicing distinctions.

5.5 Discussion of Results

This chapter has presented results from perceptual studies as well as machine recognition results on the same task. Performance of humans at recognizing words in whisper appear much better when compared with results on nonsense CV identification. There is strong evidence here that context effects aid communication, especially in whisper. The distinctive feature that is most affected in whisper is voicing, and our studies with large text corpora seem to indicate that perhaps the overall contribution of voicing distinct phonemes

Table 5.15: Estimate of information carried by voicing in speech.

	Model Cross Entropy			Information Transmitted in bits per token		
	1-grams	2-grams	3-grams	1-grams	2-grams	3-grams
baseline	1.5095	1.3063	1.2374	-	-	-
/b/ vs /p/	1.4925	1.2910	1.2241	0.0169 (1.1211%)	0.0153 (1.1708%)	0.0133 (1.0766%)
/d/ vs /t/	1.4615	1.2657	1.2014	0.0480 (3.1786%)	0.0406 (3.1046%)	0.0361 (2.9142%)
/g/ vs /k/	1.4962	1.2949	1.2274	0.0132 (0.8772%)	0.0114 (0.8725%)	0.0100 (0.8091%)
/f/ vs /v/	1.4927	1.2941	1.2283	0.0167 (1.1079%)	0.0121 (0.9294%)	0.0092 (0.7408%)
/s/ vs /z/	1.4783	1.2820	1.2169	0.0312 (2.0648%)	0.0243 (1.8605%)	0.0205 (1.6541%)
/dh/ vs /th/	1.5002	1.2980	1.2314	0.0092 (0.6120%)	0.0083 (0.6355%)	0.0061 (0.4898%)
/sh/ vs /zh/	1.5077	1.3046	1.2363	0.0018 (0.1163%)	0.0016 (0.1262%)	0.0011 (0.0928%)
/ch/ vs /jh/	1.5048	1.3019	1.2338	0.0047 (0.3110%)	0.0044 (0.3353%)	0.0036 (0.2942%)
voicing/plosives	1.4313	1.2388	1.1776	0.0782 (5.1785%)	0.0674 (5.1622%)	0.0598 (4.8318%)
fricatives/affricates	1.4458	1.2554	1.1965	0.0636 (4.2146%)	0.0508 (3.8926%)	0.0410 (3.3100%)
`[voiced]` vs `[voiceless]`	1.3677	1.1843	1.1318	0.1418 (9.3923%)	0.1220 (9.3406%)	0.1057 (8.5382%)

to communication is not large. Finally, our results, taken in conjunction with prior work, indicate that whispering works really well under noise-free listening conditions. With the exception of voicing, there is not much degradation in the transmission of distinctive features. However, it is not completely clear how well whisper works under noise. A possible follow-up study would be to study the confusions of whispered nonsense CV syllables under differing noise conditions, repeating the work of Miller and Nicely [77] but for whispered speech under different band-pass conditions.

CHAPTER 6

RECOGNITION OF WHISPERED SPEECH

This chapter describes the experiments done on recognizing whispered speech. We document the performance of standard training and adaptation techniques, and propose new methods for building whispered speech acoustic models.

6.1 Accuracy of Speaker-Independent Acoustic Models

We first took a look at the performance of speaker-independent speech recognition systems on whispered speech. Acoustic models were trained from various subsets of the data using a training recipe similar to the one found in [143]. The front end used 13 mel-frequency cepstral coefficients (MFCCs) and their delta and delta-deltas as feature vectors and applied cepstral mean subtraction. The acoustic model consisted of tied context-dependent triphones with mixture Gaussians for the observation probability distributions. Three-state monophone models were first trained, then short pause models inserted. Monophones were split into triphones and clustered using a decision tree. The number of Gaussians was then steadily increased and the models re-estimated. The language model was a bigram built from the existing sentences found in TIMIT, and backed off to unigram using Good-Turing smoothing. Accuracy is computed as

$$A = \frac{N - I - S - D}{N} \times 100\%, \tag{6.1}$$

where N is the number of words, and I, S and D are the numbers of insertions, substitutions and deletions in the recognized sentence, after aligning it with the reference (correct) sentence using a minimum edit distance al-

gorithm. Our baseline result with the TIMIT corpus itself was 75.6% word recognition accuracy using this procedure.

The data were divided into training, development and test subsets for each subset of a specific type of speech data. In our nomenclature the suffixes `-us` and `-sg` correspond to the parts of the corpus with a North American and a Singaporean accent respectively, and `-n` and `-w` correspond to unwhispered and whispered subsets of the data. The exact same steps were used in all cases to build the acoustic model; but depending on the training data-set, different acoustic models were built. The development sets were used to tune parameters such as word insertion penalty and grammar scale factor. Each acoustic model was then cross-tested with test subsets of the respective type of speech data, to obtain the results shown in Table 6.1.

The trained models show relatively high accuracy in cases where there is no mismatch between test and training data; however, the models for whisper perform significantly worse than those for non-whispered speech models. The results when there is a mismatch in the training and test data seem to indicate that accent causes at least as much accuracy loss (if not more) as speaking style when it comes to speech recognition. Poor performance across speaking style underscores the inherent brittleness in the standard approach to training speech recognizers. These figures sharply contradict those reported in [69], which claimed that whispered models can work for non-whispered speech and vice versa. There can be many reasons why our system does not perform as robustly as those trained by Itoh et al., one reason could be the different choice of front end, perhaps specifically in the cepstral normalization technique. Due to the formula used for computing accuracy, negative values are possible and are in fact reported here. We found that the errors were largely due to spurious insertions which drastically reduced accuracy. However the source of these errors remains unidentified. Very likely, techniques similar to those used for making robust speech recognizers have to be used to achieve similar performance. We now turn to a consideration of such methods.

Table 6.1: Word recognition accuracy across different models and data-sets.

(a) TIMIT and coarser wTIMIT subsets

Acoustic Models	Dataset					
	TM	wTM	wTM-n	wTM-w	wTM-us	wTM-sg
TIMIT	75.57	-17.49	-15.34	-19.63	-11.23	-21.63
wTIMIT	-9.19	81.76	87.09	76.46	84.93	79.66
wTIMIT-n	-1.02	51.19	85.99	16.74	56.58	47.61
wTIMIT-w	-6.34	50.30	25.42	76.18	53.40	48.25
wTIMIT-us	-10.22	58.34	48.75	82.24	86.27	12.95
wTIMIT-sg	-3.22	27.08	31.64	22.22	17.55	80.76

(b) TIMIT and finer wTIMIT subsets

Acoustic Models	Dataset					
	us	us-n	us-w	sg	sg-n	sg-w
TIMIT	-11.23	-10.78	-11.69	-21.63	-18.36	-24.91
wTIMIT	84.93	88.69	81.16	79.66	86.03	73.35
wTIMIT-n	56.58	88.89	24.05	47.61	84.07	11.86
wTIMIT-w	53.40	27.79	81.36	48.25	23.85	72.76
wTIMIT-us	86.27	89.41	83.12	12.95	23.39	2.52
wTIMIT-us-n	54.33	88.84	4.67	0.04	17.07	-16.99
wTIMIT-us-w	53.80	27.88	82.04	-4.72	-16.37	6.92
wTIMIT-sg	17.55	19.25	15.86	80.76	86.96	74.56
wTIMIT-sg-n	5.08	15.08	-4.83	47.40	84.55	11.90
wTIMIT-sg-w	2.77	-11.40	16.81	48.40	22.38	74.97

6.2 Implementing Eigenvoices in HTK

Eigenvoices [161] provide a way to perform rapid adaptation by exploiting the structure in inter-speaker variation. A simple approach is to consider only the mean parameters of an acoustic model. We assume that the means for an acoustic model can be treated as a linear combination of eigenvoices. The mean parameters μ_m for each mixture can be concatenated into a giant supervector

$$\mu_k = \begin{pmatrix} \mu_{k,1} \\ \vdots \\ \mu_{k,M} \end{pmatrix}, \tag{6.2}$$

for an acoustic model of the k-th speaker with a total of M mixture Gaussians. Finding the eigenvoices involves finding spanning vectors for the sub-

space spanned by K speakers

$$S = span\{\mu_k, \mu_2, ...\mu_K\}. \tag{6.3}$$

A dimensionality reduction technique is then used to find a compact subspace of the speaker subspace, thus concentrating only on the key differences that vary from speaker to speaker. Principal components analysis (PCA) is one method whereby we can obtain a set of $\{e(1), ...e(E)\}$ eigenvectors, $E < K$, which characterize the subspace S. These eigenvectors are dubbed eigenvoices. In practice, the dimensionality of the $e(k)$ vectors is very large, so directly computing the covariance matrix of the set of speaker vectors, as is required in PCA, is intractable. One way around this is to use probabilistic principal components analysis (PPCA), which has an EM algorithm that is linear in the size of the eigenvectors and estimates the principal subspace of S [162].

The PPCA-EM algorithm does not actually produce orthogonal eigenvectors – instead it produces vectors that span a given q dimensional subspace. These vectors themselves capture inter-speaker variablity, and when used in conjunction with the mean supervector, can be used as a set of eigenvoices. Implementing eigenvoices in HTK consisted of two steps – an implementation of the PPCA-EM algorithm to produce a set of eigenvoices, and an implementation of the MLED algorithm in order to find optimal weights given the eigenvoices, for a given speaker.

6.2.1 Maximum Likelihood Eigen-Decomposition

Maximum likelihood eigen-decomposition (MLED) is a method proposed in [161], which produces a ML estimate of eigenvoice weights from given speech – the derivation of which is reproduced here. The algorithm seeks to find a model that maximizes the likelihood of the observed data

$$\hat{\lambda} \quad = \quad \arg\max_{\lambda} L(O|\lambda). \tag{6.4}$$

This is equivalent to optimizing the auxilary function in the face of unknown data ξ

$$Q(\hat{\lambda}|\lambda) \quad = \quad E[logL(O,\xi|\hat{\lambda})|O,\lambda]. \tag{6.5}$$

In the case of adapting the means for a Gaussian mixture HMM, this is equivalent to optimizing

$$Q_b(\hat{\mu},\mu) = -\frac{1}{2}L(O|\mu)\sum_{m,t}\gamma_m(t)[(o_t-\hat{\mu}_m)^T C_m^{-1}(o_t-\hat{\mu}_m)], \tag{6.6}$$

for each mixture m, where μ are mean parameters, C_m are covariance parameters, o_t are observations at time t, and $\gamma_m(t)$ is the occupation likelihood of mixture m at time t.

The model means are a linear combination of eigenvoices, given by

$$\hat{\mu}_m = \sum_{k=1}^{L} w_k e_m(k). \tag{6.7}$$

Substituting this into the auxilary function and differentiating yields the system of equations

$$\begin{aligned}\sum_{m,t}\gamma_m(t)e_m^T(k)C_m^{-1}o_t &= \sum_{m,t}\gamma_m(t)\sum_{j=1}^{L} w_j e_m^T(k)C_m^{-1}e_m(j) \\ &= \sum_{m,t}\gamma_m(t)e_m^T(k)C_m^{-1}E_m w \end{aligned} \tag{6.8}$$

for $k \in 1...L$.

Rewriting the L equations in matrix form, we obtain

$$\sum_{m,t}\gamma_m(t)E_m^T C_m^{-1}o_t \quad = \quad \sum_{m,t}\gamma_m(t)E_m^T C_m^{-1}E_m w, \tag{6.9}$$

where matrix E is an L by v matrix of eigenvoices for a mixture, where v is the dimension of the observation vector, and L is the number of eigenvoices. MLED thus can be implemented in HTK by directly accumulating the $E_m^T C_m^{-1}E_m$ and $E_m^T C_m^{-1}$ matrices, and w obtained using a linear solver.

Finally, our implementation of MLED involved the following changes to HTK:

- A new update mode (UPEIGV) to HTK
- Modification of HERest to accept the new update mode using the '-u' switch with an 'e' (for eigenvoice) flag
- Code to read in an eigenvoice matrix in binary format and an '-e' switch to HERest
- Modification to HFB.c to perform the necessary accumulation
- Code to compute w from the accumulated matrices in HERest

Experiments with the implementation showed that MLED was effective in estimating weights, and often produced a good speaker-dependent model with one iteration of the algorithm.

6.3 Speaker Adaptation with Normal and Whispered Speech

We compared the performance of different adaptation techniques for whispered speech. Results for normal speech are first presented. Table 6.2(a) shows the performance improvement of various basic techniques for normal speech, averaged over all speakers. These values are computed by subtracting the baseline accuracy of the speaker-independent model of 66.65% over the test set from the accuracy for the respective speaker adapted models. Here, MLLR(m) and MAP(m) update only the mean components; CMLLR and MAP(mv) update both means and variances. The supervectors for eigenvoice adaptation are assembled from the mean components of speaker adapted models, which themselves are either adapted with MAP(m) or MLLR(m); these correspond to the EIGV(map) and EIGV(mllr) labels.

For the MAP update methods, we found that the best results were obtained on the first iteration – further iterations tended to degrade performance. Our results show that MLLR, CMLLR and Eigenvoice using MLLR-adapted means can achieve very good adaptation results even without a lot of adaptation data. MLLR and CMLLR can improve as more adaptation data is available, but the Eigenvoice method seems to quickly hit a performance asymptote. The comparatively poorer performance of MAP could be due to

Table 6.2: Delta-word recognition accuracy – improvement over SI baseline for different speaker adaptation methods (normal acoustic models).

(a) Basic speaker adaptation methods

# of Utts	MLLR(m)	CMLLR	MAP(m)	MAP(mv)	EIGV(map)	EIGV(mllr)
5	6.67	8.52	1.43	-0.24	3.39	7.37
10	7.05	8.09	2.03	1.09	4.17	7.57
50	8.20	9.92	2.40	0.72	4.30	7.46
100	8.33	8.82	3.17	1.13	4.60	7.62

(b) Combined speaker adaptation methods

# of Utts	MLLR(m)+MAP	CMLLR+MAP	EIGV+MAP
5	5.50	0.45	-3.37
10	5.44	1.72	-3.79
50	7.52	0.77	-8.53
100	7.12	1.08	-11.97

the few number of utterances used in adaptation, as well as poor overlap of updated parameters and parameters seen in the test set. This could have an effect of overtraining the MAP models to give unsatisfactory results.

The adaptation methods can be combined to good effect. One commonly used method is to apply a coarser, more rapidly adapting technique such as MLLR, followed by a finer technique such as MAP. Results for combined approaches are further tabulated in 6.2(b). Applying MAP as a final step seemed to generally degrade performance. The reason for this is not clear, but could be due to insufficient data to properly apply MAP.

Next, we look at the same methods as applied to whispered speech. Table 6.3 shows analogous results for whispered speech. The baseline accuracy for the speaker-independent model is 54.65%. The methods involving MAP once again fail to work. Our results seem to indicate that there is a larger margin for improvement in whispered speech. It seems unlikely to explain this by suggesting that there is greater inter-speaker variability in whispered speech. Another more likely explanation is that the poorer performance of the speaker-independent whispered speech model allows a larger margin for improvement. To further examine these claims we tabulate the relative improvement of each method – that is improvement in word error rate divided by the word error rate of the speaker-independent model – to give a better comparison in Table 6.4. These figures clearly support the idea that the speaker differences in whisper are greater than in normal speech.

Table 6.3: Delta word recognition accuracy – improvement over SI baseline for different speaker adaptation methods (whispered acoustic models).

(a) Basic speaker adaptation methods

	MLLR(m)	CMLLR	MAP(m)	MAP(mv)	EIGV(map)	EIGV(mllr)
5	7.60	7.94	2.46	-0.37	7.90	11.37
10	10.12	9.64	2.99	0.26	8.04	11.80
50	12.42	12.54	2.70	1.36	8.44	11.97
100	12.40	13.45	3.50	1.74	8.22	11.93

(b) Combined speaker adaptation methods

	MLLR(m)+MAP	CMLLR+MAP	EIGV+MAP
5	5.62	-0.03	-1.38
10	8.72	0.80	-3.78
50	10.05	0.67	-10.99
100	10.43	2.00	-13.82

Table 6.4: Relative WER reduction – comparison of speaker adaptation methods (in %).

	# of Utterances							
Adaptation	Normal Acoustic Model				Whisper Acoustic Model			
Method	5	10	50	100	5	10	50	100
MLLR(m)	17.88	19.13	22.37	22.89	14.62	20.30	25.23	25.29
CMLLR	22.27	20.74	25.85	23.40	15.57	18.83	25.58	26.45
MAP(m)	0.26	1.56	-1.32	0.88	1.73	1.99	-1.79	1.38
MAP(mv)	-11.55	-6.70	-15.64	-15.60	-6.00	-8.55	-9.29	-6.96
EIGV(map)	8.70	10.38	10.50	11.24	17.84	17.52	19.06	18.47
EIGV(MLLR)	20.63	20.93	20.28	20.73	25.48	25.69	26.14	26.10

6.4 Adapting Speaking Style and Accent Using CMLLR

The next set of experiments use constrained maximum likelihood linear regression (CMLLR) to determine how well existing trained non-whispered speech or whispered speech models can be adapted to a different speaking style. The best combination of parameters found was to use CMLLR with a large number of nodes (256) in our regression tree. Improvement achieved by adapting the non-whispered speech model for North American English to whispered speech is shown in the first line in Table 6.5, and further results for different permutations follow. The last column of the table corresponds to accuracy obtained from testing the target adaptation test data with a

Table 6.5: Word recognition accuracies obtained from adapting accent and speaking style with CMLLR.

Original Model	Adaptation Data (Target)	Number of Utterances 10	50	100	Target Model
wtimit-us-n	whisper	37.5	51.7	55.7	76.3
	accent	13.4	30.9	36.5	82.5
wtimit-us-w	normal	64.2	73.8	75.4	87.7
	accent	0.5	12.9	17.8	69.0

speaker-independent model trained from the training subset of the same type of data. Use of more than 100 utterances did not noticeably improve the result, and we can see that the performance never reaches the level achieved by acoustic models in the same modality. The approach works well for adapting whispered speech models to non-whispered speech but not for accent. Even for the purpose of style adaptation there is room for improvement, and other methods for adaptation should prove useful to this problem.

6.5 Acoustic Model Adaptation with Limited Whisper Data

In real-world applications it is far more likely to encounter collections of normally spoken, unwhispered speech for a particular talker rather than whispered speech. One application of interest is thus to find some method of building a speaker-dependent acoustic model of whisper using only unwhispered speech from that speaker in conjunction with whispered and unwhispered speech from anybody else. In this section, we develop some new algorithms to do just that.

6.5.1 Mapping of Eigenvoice Weights

Our algorithms are based on the eigenvoice adaptation of the means. For speaker i, the mean parameters for a given state s and mixture j are expressed as a linear combination of eigenvoices,

$$\hat{\mu}_i^{s,j} = \sum_k w_{i,k} e_k^{s,j}, \tag{6.10}$$

where e_k is a supervector made from concatenating $e_k^{s,j}$ in a particular order of the state and components, and e_k is the k-th eigenvoice for the given subspace, and $w_{i,k}$'s are eigenvoice weights that characterize the specific speaker i in the inter-speaker space. Our training procedure pairs up speaker-dependent models of unwhispered and whispered speech for individual talkers, and builds two eigenspaces – one corresponding to unwhispered speech, the other to whispered speech. This general approach is illustrated in Figure 6.1. For each speaker in the training database, the eigenvoice weights for the non-whispered and whispered subspaces are functionally related via

$$\begin{aligned} \hat{w_i}^{(w)} = (\hat{w}_{i,1}^{(w)}, ...\hat{w}_{i,k}^{(w)}, ...)^T &= f((w_{i,1}^{(n)}, ...w_{i,k}^{(n)}, ...)^T) \\ &= f(w_i^{(n)}). \end{aligned} \tag{6.11}$$

Here, the superscripts n and w correspond to unwhispered and whispered models, and f is some function between the normal and whisper spaces that has to be learned. Given a new speaker, unwhispered speech can be used to obtain suitable eigenvoice weights which are then mapped into whispered speech, and from here used to generate an acoustic model for recognizing whisper. This procedure does not require any whispered speech from the new speaker at all.

Two methods were considered for finding the mapping f. The first approach is to simply consider a least squares projection on the set of paired vectors for speakers in the training database. Let us consider a projection P that maps a vector of normal speaker weights $w_k^{(n)}$ to a vector of whispered speaker weights $w_k^{(w)}$, for the k-th speaker. A suitable error criterion is to minimize the mean squared error over all speakers, that is

$$\begin{aligned} P &= \arg\min_P \sum_{\forall k} \|Pw_k^{(n)} - w_k^{(w)}\|_2 \\ &= \arg\min_P \|{W^{(n)}}^T P^T - {W^{(w)}}^T\|_2, \end{aligned} \tag{6.12}$$

where $W^{(n)}$ are vectors of each speaker's eigenvoice weights for the unwhispered speech, arranged column by column, and $W^{(w)}$ arranged from eigenvoice weights for whispered speech. Equation 6.12 is a standard least squares problem and is solved with the pseudo-inverse of ${W^{(n)}}^T$. For K eigenvoice

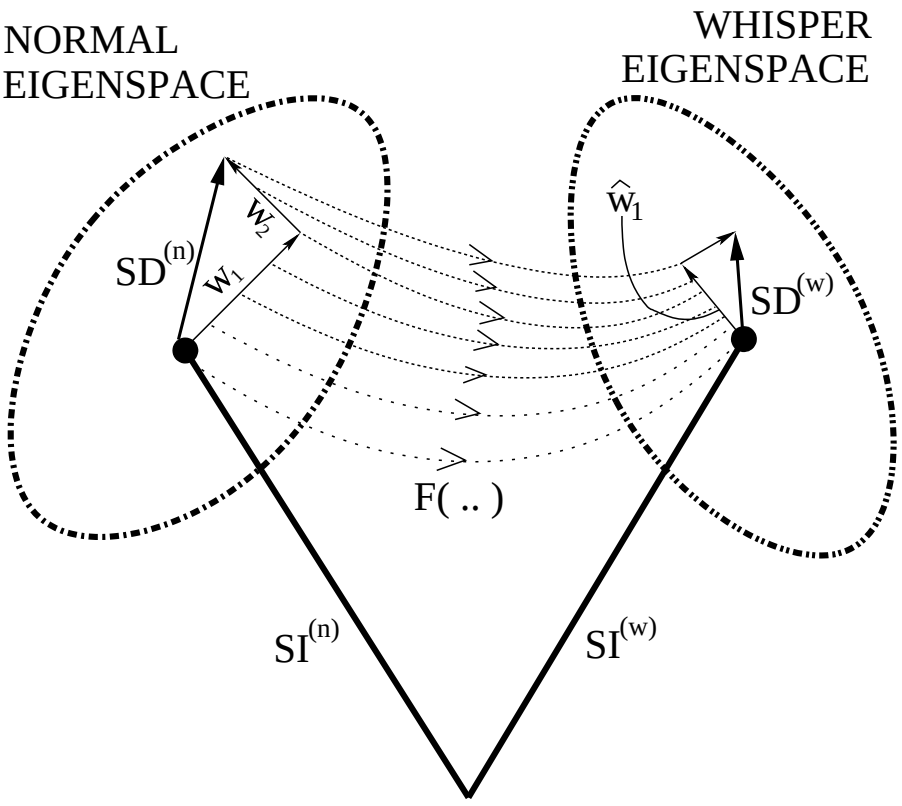


Figure 6.1: Illustration of eigenspace mapping approach. This shows the case for two eigenvoices. Thick solid lines with the bullet end represent the mean vector for speaker-independent acoustic models. Ellipses represent eigenspaces for both types of speech: the longer axis of the ellipse is aligned with the first eigenvoice, the lateral axis with the second eigenvoice. The thick arrowed lines represent the speaker-dependent perturbation off the SI-mean, which can be described by eigenvoice weights (w's). As described in the text, any plausible mapping function described by Equation 6.11 will work.

weights, this gives

$$\begin{aligned} P &= \begin{pmatrix} p_{0,0} & p_{1,0} & \cdots & p_{K-1,0} \\ p_{1,0} & p_{1,1} & \cdots & p_{K-1,1} \\ \vdots & \vdots & & \vdots \\ p_{K-1,0} & p_{K-1,1} & \cdots & p_{K-1,K-1} \end{pmatrix} \\ &= [(W^{(n)}W^{(n)^T})^{-1}W^{(n)}W^{(w)^T}]^T. \end{aligned} \tag{6.13}$$

Since the eigenvoice supervectors are themselves orthonormal, a linear transform with them as column vectors is distance-preserving. Hence our so-

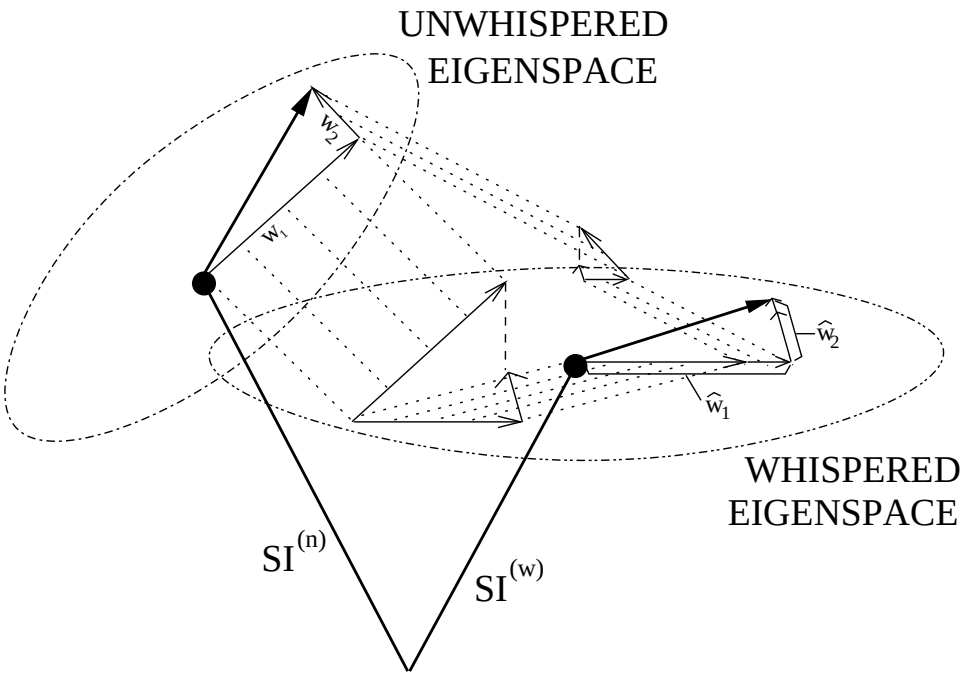


Figure 6.2: Illustration of projection-based eigenvoice mapping procedure. This shows the case for two eigenvoices. Thick solid lines with a bullet end represent the mean vector for speaker-independent acoustic models. Ellipses represent eigenvoice spaces for both types of speech: the longer axis of the ellipse is aligned with the first eigenvoice, and the lateral axis with the second eigenvoice. The thick arrowed line on the left represents the speaker-dependent perturbation off the SI-mean (left arrow); the other arrowed line corresponds to the resulting mapped SD-perturbation. We can see that the estimates of eigenvoice weights in whisper space $\hat{w}_k$'s result from a linear projection of w_k's.

lution also minimizes $\|E^{(w)}W^{(w)} - E^{(w)}PW^{(n)}\|_2 = \|\mu^{(w)} - \mu^{(w)}_{E^{(w)\perp}} - E^{(w)}PW^{(n)}\|_2$, where $E^{(w)}$ is a matrix composed of the whispered eigenvoices, and $\mu^{(w)}$ is the supervector of Gaussian mean parameters of a particular speaker's whispered acoustic model. Since $E^{(w)}W^{(w)}$ is an approximation to $\mu^{(w)}$, leaving out the components of $\mu^{(w)}$ which are orthogonal to the $E^{(w)}$ subspace, this has to be accounted for by $\mu^{(w)}_{E^{(w)\perp}}$. In other words this approach approximately minimizes the mean squared error between the resulting mean parameters of the projected supervector computed from parameters of the original normal acoustic model and the original whispered acoustic model. This approach is illustrated in Figure 6.2.

A second method models the non-whispered and whispered eigenvoice weights $w_i^{(n)}$ and $w_i^{(w)}$ of each speaker i jointly using a Gaussian mixture

PDF, each mixture having a mean and covariance

$$\lambda_k = \begin{pmatrix} \lambda_k^{(n)} \\ \lambda_k^{(w)} \end{pmatrix}^T, \tag{6.14}$$

$$\Sigma_k = \begin{pmatrix} \Sigma_k^{(n)} & \Sigma_k^{(nw)} \\ \Sigma_k^{(wn)} & \Sigma_k^{(w)} \end{pmatrix}, \tag{6.15}$$

and applying Gaussian mixture regression [4] to obtain an estimate of $\hat{w}^{(w)}$,

$$\hat{w}_i^{(w)} = \sum_k p(k|w^{(n)}) C_k (w^{(n)} - \lambda_k^{(n)}), \tag{6.16}$$

where C_k is the rotation matrix for the k'th mixture given by

$$C_k = \Sigma_k^{(wn)} \Sigma_k^{n-1}. \tag{6.17}$$

In practice, these methods may provide a poor estimate of the whispered speech supervector due to insufficient data. One way to ameliorate this is to use the speaker-independent whispered speech model $\mu_{SI}^{(w)}$ as a background and perform a MAP update on each state j and mixture component m with

$$\tilde{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau} \hat{\mu}^{(w)} + \frac{\tau}{N_{jm} + \tau} \mu_{SI}^{(w)}, \tag{6.18}$$

where N_{jm} are state occupation likelihoods.

6.5.2 Joint Eigenvoice Adaptation of Whispered and Normal Speech

In this approach we train common eigenvoices as an alternative to using a mapping function. We can obtain supervectors for the speaker-dependent normal speech model and the whispered speech model, and concatenate them into a giant supervector

$$\tilde{\mu}_i = \begin{pmatrix} \mu_i^{(n)} \\ \mu_i^{(w)} \end{pmatrix}, \tag{6.19}$$

where $\mu^{(n)}$ and $\mu^{(w)}$ are supervectors of speaker i, described in the previous section, for normal and whispered speech respectively. PPCA is then used

to find a set of "joint" eigenvoices $\tilde{e}$'s from

$$\begin{aligned} SVD\{\begin{pmatrix}\tilde{\mu}_1 & \tilde{\mu}_i & ... & \tilde{\mu}_K\end{pmatrix}\} &= \tilde{E}\Omega V \\ &= \begin{pmatrix}\tilde{e}_0 & \tilde{e}_1 & ... & \tilde{e}_M\end{pmatrix}\Omega V. \\ &= \begin{pmatrix} e_0^{(n)} & e_1^{(n)} & ... & e_M^{(n)} \\ e_0^{(w)} & e_1^{(w)} & ... & e_M^{(w)}\end{pmatrix}\Omega V. \end{aligned} \quad (6.20)$$

Given normal speech from a new speaker, we simply use MLED to estimate weights using only the top half of the eigenvoices, and apply the same weights in conjunction with the whisper halves of the eigenvoices to generate a whispered speech model.

6.5.3 Word Recognition Accuracy Achieved Using Transformed Models

For this set of experiments, a triphone model with single Gaussians at each state was used, so that memory requirements for eigenvoice computation were tractable. Only the subset of speakers with a North American Accent (`wtimit-us`) was used. Our previous experiment found that eigenvoice adaptation worked better using MLLR-based speaker adapted models as opposed to MAP adapted ones when building the subspaces, so MLLR-based eigenvoices were used for experiments here. One speaker is taken out for testing, and the remaining used in the training procedures outlined in section 6.5. The results are cross-validated across all 28 speakers, and averaged accuracy is shown in Table 6.6. The methods labeled `MAP` are identical to the two methods, except that instead of using the synthesized whispered speech acoustic model directly, an additional MAP adaptation step was used to adapt from the speaker-independent whisper model towards them. The label "Shared Wts" refers to the joint eigenspace approach outlined in section 6.5.2. The baseline accuracy for using the speaker-independent whispered acoustic model was 57.54%. The NIST `sctk` toolkit was used to compare the recognition results from the speaker-independent acoustic model and those from the various systems, and all differences were found to be statistically significant.

Our results seem to indicate that the effectiveness of the additional MAP

Table 6.6: Word recognition accuracies of speaker-dependent whisper acoustic model produced from normal speech of said talker.

	Number of Utterances				
	5	10	50	100	200
Baseline (SI-Whisper)	57.54%				
Lin Proj.	55.28	56.41	56.56	56.45	56.28
GMM	62.42	62.06	62.12	62.50	62.58
Proj+MAP	58.14	58.36	58.18	58.40	57.97
GMM+MAP	57.99	58.15	58.18	58.44	58.03
Shared Wts	66.39	66.50	66.32	66.48	66.57

adaptation step depends on the original effectiveness of the algorithm. For instance, the linear projection algorithm produces whisper models that fare worse than the speaker-independent model. Adding the MAP step gives positive improvements to this technique. The GMM approach produces models that work better, and applying MAP reduces its effectiveness. The most effective method thus far is to treat the whisper and normal subspaces together and allow PPCA to derive a set of joint eigenvoices. It is not clear why this method works better, and more study into variations of these algorithms seems necessary.

6.6 Summary

This chapter has documented numerous experiments with basic speech recognition and various adaptation methods on normal and whispered speech. Speaker-independent acoustic models trained on whispered speech data perform favorably compared with those trained on unwhispered speech data. The experiments using different adaptation techniques illustrate the comparative performance of various speaker adaptation techniques on whispered speech data. Speaker adaptation has a greater impact in whispered speech than unwhispered speech, though the reason for this is not clear.

Adapting unwhispered acoustic models with whispered speech results in an acoustic model for whispered speech that generally works well, but the resulting model does not perform better than a speaker-independent whispered speech acoustic model. We also develop three algorithms to speaker-adapt a whispered acoustic model using *unwhispered* speech data. Of these

approaches, characterizing whispered and unwhispered acoustic model parameters in a joint subspace works the best. A promising avenue for further work seems to be along the line of such algorithms.

CHAPTER 7

CONCLUSION

7.1 Summary of Completed Work

So far we have presented two sets of experiments, perceptual and simulation-based, to shed some light on whispered speech and how it is recognized. Our perceptual results are consistent with prior work in showing that whisper carries much information, and despite first impressions, phonemic voicing is not all lost in whisper. More importantly, we can quantify how well whispering actually works at word-level contexts. A simple experiment with context-length shows that perhaps there is actually not much information carried by phonemic voicing, and that contextual information which greatly aids communication is emphasized even more in whisper.

Our experiments with speech recognition algorithms seem to suggest that the standard approach is not foolproof. Despite claims by Itoh et al., a normal speech acoustic model does not do well at recognizing whisper; adaptation methods need to be used. We consider the problem of performing recognition with limited amounts of training whispered speech, and propose a novel algorithm to do so based on eigenvoices.

However there are many problems left unsolved. Even as we know whisper conveys information relatively well, precisely what acoustic correlates help it to do so remain unknown. Furthermore, the many problems associated with speech technology, in speaker identification, recognition, and understanding have their counterparts in whisper, and these problems have to be further worked on separately. We now expand on some of these problems and highlight some possible avenues for future work.

7.2 Future Work

7.2.1 Expansion of the wTIMIT Corpus

One drawback of the wTIMIT corpus is the relatively small number of speakers. TIMIT for instance has over 630 speakers. A third stage of collection involving fewer sentences per speaker might be beneficial to the corpus. With this addition, experiments involving speaker identification and gender classification become meaningful to do.

7.2.2 Hyperarticulation in Whisper

The literature [63] seems to suggest that there is some hyperarticulation going on in whisper. Although some studies have been made, to date there has been no study of the movement of the tongue. In whisper, it appears that articulators move to preserve salient acoustic targets [53]; one question to investigate would be the nature of hyperarticulation and what acoustic cues it enhances. In some sense, one might wonder if a "clear speech" effect exists [173] for whisper – speakers could be modifying acoustic characteristics such as speaking rate, number and duration of pauses, or making small but beneficial phonological substitutions [174], to ameliorate the effects of a channel perceived to be of otherwise less intelligibility. Learning about this could let us better understand what acoustic cues are important for perception of different phonemes. One promising approach would be to collect articulation data from speakers during whispering, and compare it to when phonated speech is used. One method of doing so would be to use an electromagnetic midsagittal articulometer [175] (EMMA) or similar system to track the movement of the tongue.

7.2.3 Discovering a Common Phonetic Process of Whisper and Phonated Speech

Alan Poritz [176] performed an experiment in which he found that ergodic hidden Markov models using LPCs as the feature vectors managed to discover vowel and consonant structure in running speech. The question of whether

or not such a method would work for whispered speech is intriguing – in whisper, voicing cannot allow an easy segregation of phoneme classes.

An initial experiment in this vein has been considered, but results are inconclusive. We considered the continuous variable duration hidden Markov model [177], shown in Figure 7.1(a). In this ergodic model, the duration of a speech segment is explicitly modeled using the gamma distribution, and observations are modeled with a multivariate Gaussian probability distribution. Our initial experiment to discover speech segments with MFCC vectors from speech did not return labels that corresponded to phonemically distinguishable segment types. More work needs to be done – also there is a further experiment which appears to be even more interesting.

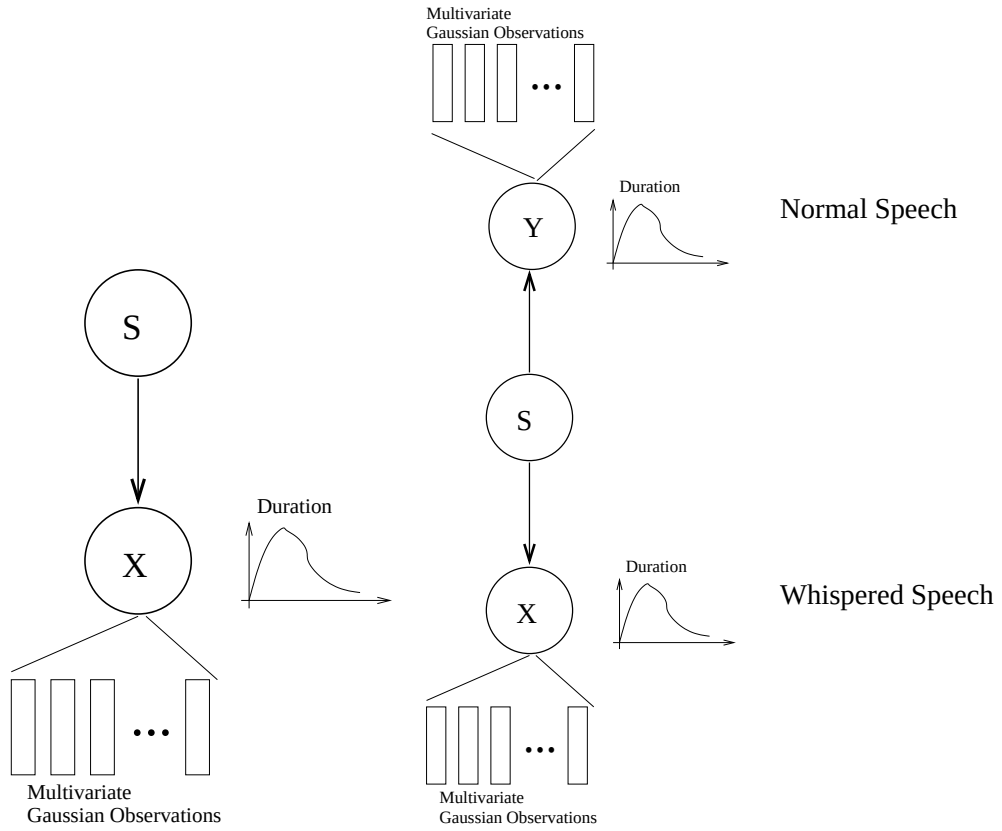
We may want to consider the so-called "multistream" version, shown in Figure 7.1(b). The motivation for this is to assume a single underlying phonetic process that can generate both normal and whispered speech. Such a model would produce asynchronous streams of observations, each stream having its own duration and observation model, which are conditionally independent of the other given the state. In working out the mathematics of EM, it appears that the same update equations can be used, except for the state transitions, which are just averages of the individual per-stream computed updates. What such an ergodic system would discover will have to be left as a future experiment.

7.2.4 Modeling the Whispered Glottal Source as Noise

It is commonly mentioned that in whisper the glottal excitation is somewhat noise-like. It is important to confirm if this is actually true. One approach would be to examine the LPC spectra of whispered speech for different phonemes, and look for the poles near the low frequency regions, which do not correspond to any formants. In this way we can test if whisper is indeed excited by spectrally shaped noise.

7.2.5 Verification of the Stevens-Wickesberg Result

Wickesberg and Stevens played whispered consonants /t/ and /d/ to chinchillas and recorded the auditory response [72]. They found that the response



(a) Single stream (b) Multi-stream

Figure 7.1: Continuous variable duration hidden Markov models.

to /da/ had a double onset as opposed to a single onset in /t/. It is not known whether an auditory front end would produce the same result, and this is worth investigating in detail.

7.2.6 Further Algorithms for Adaptation

In this work we have outlined some new algorithms for adapting normal speech acoustic models to whisper. However, our method still requires some whispered speech for initial training. The question as to how little whispered speech we can work with is up for investigation, and new algorithms remain to be discovered.

7.2.7 Isolating the Critical Points in Whispered Speech

Furui [178] performed perceptual experiments involving front and back-truncated consonant vowels in order to isolate where perceptually important acoustic information in the Japanese syllable is found. He discovered a perceptual critical point where the identification of the truncated syllable changes rapidly as a position of function, which is related to the position of maximum spectral transition. He concluded that the position of maximum spectral transition contained the most important information for both consonant and vowel identification. Whether this is so for whispered speech is unclear, and is worth investigating. Rather than conduct a perceptual experiment, an alternative is to use machine recognition, especially a universal classifier such as support vector machines [179] to work on features extracted from truncated CV sequences. The identification rates can be plotted along the timeline and the critical point identified. In this way we can confirm or refute Furui's result for whispered speech, and this would be critical to understanding what acoustic cues are invariant for phoneme perception.

7.3 Conclusion

In the end, this work only scratches the surface of a very deep and involved problem. We leave the reader with numerous possible future directions to take with this research, in the hope that exciting discoveries will await us.

APPENDIX A

FORMANT MEASUREMENTS

This appendix has formant measurements of the vowels /a,i,u/ in fluent speech. The vowels are taken from the following contexts:

- /a/ is taken from "power" ([p a w ɚ])
- /i/ is taken from "rarely" ([r e r l i])
- /u/ is taken from "hindu" ([h i n d u])

The utterances used were:

- A huge power outage rarely occurs.
- Does Hindu ideology honor cows?

Table A.1 show such measurements for made on the wTIMIT-us sub-corpus, and Table A.2 for the wTIMIT-sg sub-corpus.

Table A.1: Formant frequencies for /i,a,u/ in fluent speech (wTIMIT-US).

	Normal Speech									Whispered Speech								
	/a/			/i/			/u/			/a/			/i/			/u/		
Speaker	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
s101	850	1510	2530	425	2227	3004	492	1462	2573	1133	1162	2731	765	2381	3015	552	1563	2692
s102	506	1216	2350	409	2685	3184	404	1585	3318	1056	1743	3002	769	2263	3069	470	1713	2889
s103	771	1203	2321	335	1936	2566	372	1208	2300	920	1249	2224	561	1923	2768	736	1290	2341
s105	928	1620	2669	418	2457	2985	492	1529	2691	1025	1710	2947	478	2525	2836	700	1400	2857
s106	817	1607	2444	319	2218	2517	269	884	2465	928	1789	2622	543	2343	2597	404	1054	2913
s107	771	1134	2255	400	1871	2520	386	1308	2286	974	1603	2212	454	2060	2708	441	1308	2488
s108	945	1584	2465	401	2230	2881	441	1677	2913	1141	1769	2676	938	2320	2926	884	1788	3152
s109	934	1703	2719	438	2291	3198	404	1474	2876	1041	1694	2890	644	2544	3134	607	1493	2531
s111	627	1146	2157	326	2159	2454	331	1218	2534	1099	1304	2358	566	1910	2303	672	1128	2327
s112	788	1394	2594	406	2498	3282	546	1313	2992	1051	1592	2848	600	2436	3041	878	1349	2927
s116	904	1569	2363	394	2292	2535	456	1239	2616	1141	1810	2712	518	2443	2855	671	1241	2750
s117	799	1238	2274	371	2052	2620	401	1430	2610	1047	1639	2781	462	2190	2721	601	1621	2643
s118	782	1246	2236	290	2152	2828	370	1150	2712	951	1533	2615	552	2801	3617	698	1290	2691
s119	893	1414	2451	485	2056	2666	390	1163	2245	1032	1640	2691	960	2186	2900	644	1144	2373
s120	1037	1479	2750	414	2318	2773	400	1252	2528	1133	1967	2685	408	2413	2706	733	1427	2625
s121	715	1255	1777	392	1886	2633	362	1077	2587	1147	1940	2636	393	2064	2707	533	1234	2488
s122	1181	2251	3556	553	1511	2359	411	1204	2470	976	2304	3337	408	2051	2714	497	1161	2562
s123	989	1433	2407	405	2270	2818	376	1171	2820	1070	1770	2313	1050	2857	3508	515	1382	2673
s124	700	1271	2322	359	2012	2536	362	1400	2328	958	1714	2691	760	2190	2794	736	1474	2562
s125	792	1315	3130	443	2255	3164	506	1529	3374	1117	1568	2913	1066	2379	3233	948	1647	3080
s128	738	1331	2354	406	1805	2617	440	1147	2564	847	1493	2544	510	1905	2709	570	1216	2433
s129	884	1439	2683	405	2350	3135	424	1120	2777	958	1269	2587	511	2554	3287	662	1216	2839
s130	902	1321	2142	427	2180	2594	476	1270	2681	1227	1961	2746	842	2186	2827	884	1290	2629
s131	738	1228	2420	402	1864	3221	446	1290	2286	939	1529	2525	607	1759	2508	552	1198	2636

Table A.2: Formant frequencies for /i,a,u/ in fluent speech (wTIMIT-SG).

	Normal Speech									Whispered Speech								
	/a/			/i/			/u/			/a/			/i/			/u/		
Speaker	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
s000	815	1280	2817	278	2599	3189	317	1243	2672	1084	1633	2593	462	2539	3113	623	1176	2929
s001	956	1204	3138	364	2659	3147	378	925	2863	1120	1599	3212	1326	2655	2960	750	1142	2876
s002	737	1135	2281	324	2361	2553	257	920	2618	975	1294	2133	1527	2718	3715	441	1080	2766
s003	752	1107	2751	243	2328	3289	337	857	3157	803	1202	2741	399	2515	2544	399	974	2798
s004	581	1064	2590	262	2240	3019	311	867	2572	828	1016	2641	367	2199	2968	549	1198	2415
s005	667	1039	2776	290	2167	3282	485	1174	2831	1166	1341	2463	633	2309	2928	514	1234	2868
s006	799	1352	2995	284	2768	3859	423	1081	3160	1151	1756	2940	316	2698	3663	653	1032	2359
s007	740	955	2675	291	1972	3137	394	1077	2506	1182	2360	3575	404	2202	3171	998	2361	3245
s008	868	1301	2904	385	2577	3371	371	1271	3108	1111	1354	2842	573	2529	3209	786	1284	3021
s009	764	1278	1831	363	2792	3388	524	992	3373	1119	1546	2711	788	1981	2893	-	-	-
s010	620	1014	2296	280	2073	2911	339	877	2506	1015	2118	3416	386	2064	3023	-	-	-
s011	717	1097	2435	301	2138	2797	379	1131	2946	928	1107	2527	564	2114	3145	958	2488	3466
s012	586	1067	2839	297	2341	2949	342	1563	2563	1212	2382	3226	321	2108	3019	490	1327	2581
s013	519	956	3097	473	2741	3310	381	1364	2903	1081	1602	3145	902	2802	3374	679	1474	3023
s014	550	1056	2677	262	2097	3054	328	980	2721	832	1077	2845	356	2089	2907	405	1364	2359
s015	579	1025	2464	307	2428	3107	330	953	2660	990	1449	2796	330	2160	2950	773	1290	2765
s016	660	1148	2672	302	2307	3329	354	1189	2835	956	1255	2602	444	2488	3563	552	1179	3060
s017	786	1180	2460	402	2194	2955	393	1074	2646	1005	1412	2651	691	2194	2505	821	1178	3072
s018	700	1105	2581	343	2662	3643	478	1185	3159	1327	2654	3706	774	2673	3576	607	1234	2618
s019	750	1320	2823	339	2762	3598	371	1172	3317	1025	1578	2640	591	2584	3374	847	1234	3042

REFERENCES

[1] W. H. Perkins and R. D. Kent, *Functional Anatomy of Speech, Language and Hearing: A Primer.* San Diego, CA: College-Hill Press, Inc., 1986.

[2] E. L. DuBrul, *Evolution of the Speech Apparatus.* Springfield, IL: American Lecture Series. Monograph in American Lectures in Anatomy, 1958.

[3] M. H. Johnson, Nov. 2010, private correspondence.

[4] A. Acero, X. Huang, and H. Hsiao-Weun, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Upper Saddle River, NJ: Prentice-Hall, 2001.

[5] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics 30. Cambridge, MA: MIT Press, 1999. [Online]. Available: http://linguistlist.org/pubs/books/get-book.cfm?BookID=540

[6] P. Lieberman and S. E. Blumstein, *Speech Physiology, Speech Perception and Acoustic Phonetics.* Cambridge, MA: Cambridge University Press, 1988.

[7] J. B. Allen, *Articulation and Intelligibility.* San Rafael, CA: Morgan and Claypool, 2005.

[8] S. E. Levinson, *Mathematical Models for Speech Technology.* Hoboken, NJ: John Wiley and Sons, May 2005.

[9] B. Sklar, "Defining, designing, and evaluating digital communication systems," *IEEE Communications Magazine*, vol. 31, no. 11, pp. 91–101, November 1993.

[10] P. Ladefoged, *A Course in Phonetics.* Fort Worth, TX: Harcourt Brace Jovanovich College Publishers, 1993.

[11] J. Kramsky, *The Phoneme: Introduction to the History and Theories of a Concept.* München, Germany: International Library of General Linguistics, 1974.

[12] Carnegie Mellon University, "The CMU pronouncing dictionary," 2010. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[13] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, UK: Cambridge University Press, June 1999.

[14] P. Roach, *English Phonetics and Phonology*. Cambridge, UK: Cambridge University Press, 1983.

[15] P. V. Reenen, *Phonetic Feature Definitions: Their Integration into Phonology and Their Relation to Speech: A Case Study of the Feature NASAL*. Cinnaminson, NJ: Foris Publications, 1981.

[16] M. J. Hashimoto, "Notes on Mandarin phonology," in *Studies in General and Oriental Linguistics*, R. Jakobson and S. Kawamoto, Eds. Tokyo, Japan: TEC Co., 1970, pp. 207–220.

[17] M. Davenport and S. Hannahs, *Introducing Phonetics and Phonology*. London, UK: Arnold, a member of the Hodder Headline Group, 1998.

[18] S. E. G. Ohman, "Coarticulation in VCV utterances: Spectrographic measurements," *Journal of the Acoustical Society of America*, vol. 39, pp. 151–168, January 1966.

[19] K. Johnson, *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishing, 2004.

[20] P. Ladefoged, "A course in phonetics," Jan 2011. [Online]. Available: http://www.phonetics.ucla.edu/course/chapter1/chapter1.html

[21] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY: Harper and Row, 1968.

[22] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, January 2001.

[23] K. Hayward, *Experimental Phonetics*. Edinburgh, England: Pearson Education Limited, 2000.

[24] G. Tserdanelis and W. Y. P. Wong, Eds., *Language Files: Materials for an Introduction to Language and Linguistics*, 9th ed. Columbus, OH: Ohio State University, Department of Linguistics, 2004.

[25] J. van den Berg, "Myoelastic-aerodynamic theory of voice production," *Journal of Speech and Hearing Research*, vol. 3, no. 1, pp. 227–244, 1958.

[26] M. M. Sondhi, "Model for wave propagation in a lossy vocal tract," *Journal of the Acoustical Society of America*, vol. 55, no. 5, pp. 1070–1075, 1974.

[27] M. R. Portnoff, "A quasi-one-dimensional digital simulation for time varying vocal tract," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1973.

[28] H. Dudley, "The vocoder," *Journal of the Acoustical Society of America*, vol. 11, no. 2, p. 169, 1930.

[29] D. F. Young, B. R. Munson, and T. H. Okiishi, *A Brief Introduction to Fluid Mechanics*. New York, NY: John Wiley and Sons, 2001.

[30] A. G. Webster, "Acoustical impedence and the theory of horns and of the phonograph," in *National Academy of Science*, vol. 5, no. 7, 1919, pp. 275–282.

[31] D. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–975, 1980.

[32] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, 1971.

[33] J. Markel and A. H. Gray, *Linear Prediction of Speech*. New York, NY: Springer-Verlag, 1976.

[34] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[35] G. von Bekesy, *Experiments in Hearing*. York, PA: Maple Press Company, 1960.

[36] H. G. Nilsson and A. R. Moller, "Linear and nonlinear models of the basilar membrane motion," *Biological Cybernetics*, vol. 27, pp. 107–112, 1977.

[37] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *Journal of the Acoustical Society of America*, vol. 87, pp. 2592–2605, 1990.

[38] H. Y. Alkahby, M. A. Mahrous, and B. Mamo, "Mathematical model for the basilar membrane as a two dimensional plate," in *15th Annual Conference of Applied Mathematics*, 1999, pp. 115–124.

[39] J. B. Allen, "Cochlear micromechanics – a mechanism for transforming mechanical to neural tuning within the cochlea," *Journal of the Acoustical Society of America*, vol. 62, pp. 930–939, October 1977.

[40] R. Fettiplace and C. M. Hackney, "The sensory and motor roles of auditory hair cells," *Nature Reviews Neuroscience*, vol. 7, pp. 19–29, January 2006.

[41] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *Journal of Neurophysiology*, vol. 42, no. 4, pp. 1083–1107, July 1979.

[42] R. V. Harrison, A. Kakigi, H. Hirakawa, N. Harel, and R. J. Mount, "Tonotopic mapping in the auditory cortex of the chinchilla," *Hearing Research*, vol. 100, pp. 157–163, October 1996.

[43] B. Herrnberger, S. Kempf, and G. Ehret, "Basic maps in the auditory midbrain," *Biological Cybernetics*, vol. 87, pp. 231–240, October 2002.

[44] G. Ehret, S. Hage, M. Egorova, and B. Müller, "Auditory maps in the midbrain: The inferior colliculus," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, D. Pressnitzer, A. Cheveigné, S. McAdams, and L. Collet, Eds. New York, NY: Springer, 2005, pp. 162–168.

[45] B. Arons, "A review of the cocktail party effect," *American Journal of Voice I/O Society*, vol. 12, pp. 35–50, 1992.

[46] H. McGurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[47] A. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, pp. 358–368, 1957.

[48] S. E. Lively and D. B. Pisoni, "On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 56, pp. 1665–1679, 1997.

[49] R. P. Lippman, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[50] M. D. Riley, *Speech Time-Frequency Representations.* Boston, MA: Kluwer Academic Publishers, 1989.

[51] L. F. Lamel, "Formalizing knowledge used in spectrogram reading: Acoustic and perceptual evidence from stops," Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, MA, Tech. Rep. 367, December 1988.

[52] C. Lashley and D. M. Hicks, "Vibratory action of the vocal folds during whisper," *The Institute for Advanced Study of the Communication Processes Bulletin*, vol. 2, pp. 32–35, 1980.

[53] T. I. P. Mills, "Speech motor control variables in the production of voicing contrasts and emphatic accent," Ph.D. dissertation, University of Edinburgh, Edinburgh, UK, August 2009.

[54] N. P. Solomon, G. N. McCall, M. W. Trosset, and W. C. Gray, "Laryngeal configuration and constriction during two types of whispering," *Journal of Speech and Hearing Research*, vol. 32, pp. 161–174, March 1989.

[55] P. Monoson and W. R. Zemlin, "Quantitative study of whisper," *Folia Phoniatrica*, vol. 36, pp. 53–65, 1984.

[56] J. Sundberg, R. Scherer, M. Hess, and F. Muller, "Whispering – a single-subject study of glottal configuration and aerodynamics," *Journal of Voice*, vol. 24, pp. 574–584, 2009.

[57] C. Zeroual, J. H. Esling, and L. Crevier-Buchman, "Physiological study of whispered speech in Moroccan Arabic," in *Interspeech*, Portugal, September 2005, pp. 1069–1072.

[58] R. J. Klich, "Effects of speech level and vowel context on intraoral air pressure in vocal and whispered speech," *Folia Phoniatrica*, vol. 34, pp. 33–40, 1982.

[59] J.-F. P. Bonnot and C. Chevrie-Muller, "Some effects of shouted and whispered conditions on temporal organization," *Journal of Phonetics*, vol. 19, pp. 473–483, 1991.

[60] E. T. Stathopoulos, J. D. Hoit, T. J. Hixon, P. J. Watson, and N. P. Solomon, "Respiratory and laryngeal function during whispering," *Journal of Speech and Hearing Research*, vol. 34, pp. 761–767, August 1991.

[61] M. F. Schwartz, "Air consumption values for oral and whispered plosive-vowel syllables," *Journal of the Acoustical Society of America*, vol. 49, no. 2B, p. 610, June 1970.

[62] M. F. Schwartz, "Bilabial closure durations for /p/, /b/, and /m/ in voiced and whispered vowel environments," *Journal of the Acoustical Society of America*, vol. 51, no. 6, pp. 2025–2029, 1972.

[63] M. Higashikawa, J. R. Green, C. A. Moore, and F. D. Minifie, "Lip kinematics for /p/ and /b/ production during whispered and voiced speech," *Folia Phoniatrica*, vol. 55, no. 1, pp. 17–27, 2003.

[64] M. F. Schwartz, "Power spectral density measurements of oral and whispered speech," *Journal of Speech and Hearing Research*, vol. 13, no. 2, pp. 445–446, 1970.

[65] K. J. Kallail and F. W. Emanuel, "Formant-frequency differences between isolated whisper and phonated vowel samples produced by adult female subjects," *Journal of Speech and Hearing Research*, vol. 27, pp. 245–251, June 1984.

[66] K. J. Kallail and F. W. Emanuel, "An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects," *Journal of Phonetics*, vol. 12, pp. 175–186, 1984.

[67] S. T. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, 2008.

[68] S. T. Jovicic, "Formant feature differences between whispered and voiced sustained vowels," *Acustica*, vol. 84, pp. 739–743, 1998.

[69] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *Acoustics Speech and Signal Processing*, vol. 1, 2002, pp. 389–392.

[70] R. W. Morris, "Enhancement and recognition of whispered speech," Ph.D. dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, August 2003.

[71] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whispered register," *Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101–2107, October 1994.

[72] H. E. Stevens and R. E. Wickesberg, "Ensemble responses of the auditory nerve to normal and whispered stop consonants," *Hearing Research*, vol. 131, pp. 47–62, May 1999.

[73] V. C. Tartter, "Identifiability of vowels and speakers from whispered syllables," *Perception and Psychophysics*, vol. 49, no. 4, pp. 365–372, 1991.

[74] G. L. Dannenbring, "Perceptual discrimination of whispered phoneme pairs," *Perceptual and Motor Skills*, vol. 51, pp. 979–985, 1980.

[75] A. S. Bregman and J. Campbell, "Primary auditory stream segregation and perception of order in rapid sequences of tones," *Journal of Experimental Psychology*, vol. 89, no. 2, pp. 244–249, 1971.

[76] V. C. Tartter, "What's in a whisper?" *Journal of the Acoustical Society of America*, vol. 86, pp. 1678–1683, November 1989.

[77] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, March 1955.

[78] M. J. Munro, "Perception of voicing in whispered stops," *Phonetica*, vol. 47, pp. 173–181, 1990.

[79] I. B. Thomas, "Perceived pitch of whispered vowels," *Journal of the Acoustical Society of America*, vol. 46, pp. 468–470, August 1969.

[80] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels," *Journal of Voice*, vol. 10, no. 2, pp. 155–158, 1996.

[81] M. Higashikawa and F. D. Minifie, "Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels," *Journal of Speech, Language and Hearing Research*, vol. 42, pp. 583–591, 1999.

[82] Y. M. Cheung, "Recognition of lexical tones in Cantonese whispered speech," M.S. thesis, The University of Hong Kong, Hong Kong, April 2003.

[83] A. S. Abramson, "Tonal experiments with whispered Thai," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, A. Valdman, Ed. The Hague, Netherlands: Mouton, 1972, pp. 31–44.

[84] M. Gao, "Tones in whispered Chinese: Articulatory features and perceptual cues," M.S. thesis, University of Victoria, British Columbia, Canada, 2002.

[85] H. Nicholson and A. Teig, "How to tell beans from farmers: Cues to the perception of pitch accent in whispered Norwegian," in *19th Scandinavian Conference of Linguistics*, vol. 31, no. 2, 2003, pp. 315–325.

[86] M. C. L. Greene and L. Mathieson, *The Voice and Its Disorders*. London, England: Whurr Publishers, 1989.

[87] N. J. Lass, K. R. Hughes, M. D. Bowyer, L. T. Waters, and V. T. Bourne, "Speaker sex identification from voiced, whispered and filtered isolated vowels," *Journal of the Acoustical Society of America*, vol. 59, pp. 675–678, March 1976.

[88] M. A. Carlin, B. Y. Smolenski, and S. J. Wendt, "Unsupervised detection of whisper speech in the presence of normal phonation," in *International Conference on Spoken Language Processing*, 2006, pp. 685–688.

[89] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, January 1983.

[90] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahamdi, "Voiced speech from whispers for post-laryngectomised patients," *International Journal of Computer Science*, vol. 36, no. 4, 2009.

[91] A. Goalic and S. Saoudi, "An intrinsically reliable and fast algorithm to compute the line spectrum pairs in low bit-rate CELP coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1995, pp. 728–731.

[92] J.-L. Gauvain and L. Lamel, "Large vocabulary continuous speech recognition: from laboratory systems towards real-world applications," *IEICE Transactions on Information*, vol. E00-1, no. 1, pp. 1–14, January 1006.

[93] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, , and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[94] J. Baker, "Stochastic modeling for speech recognition," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 1976.

[95] J. K. Baker, "The Dragon system: An overview," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-23, no. 1, pp. 24–29, February 1975.

[96] F. Jelinek, "A real-time, isolated-word speech recognition system for dictation transcription," in *International Conference on Acoustics, Speech and Signal Processing*, April 1985, pp. 858–861.

[97] G. Evermann, H. Y. Chan, M. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 17–21.

[98] K.-F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 1, pp. 35–45, January 1990.

[99] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. G. Ea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer, "The 1997 CMU Sphinx-3 English broadcast news transcription system," in *Proceedings of the 1998 DARPA Speech Recognition Workshop*, 1998, pp. 55–59.

[100] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 1691–1694.

[101] A. Lee and T. Kawahara, "Recent development of the open source speech recognition engine Julius," in *Asia-Pacific Signal and Information Processing Association*, 2009, pp. 131–137.

[102] M. Cohen, H. Murveit, J. Bernstein, P. Price, and M. Weintraub, "The Decipher speech recognition system," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 1990, pp. 77–80.

[103] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, December 2003.

[104] N. Strom, L. Hetherington, T. J. Hazen, E. Sandness, and J. Glass, "Acoustic modeling improvements in a segment-based speech recognizer," in *Proc. IEEE ASRU Workshop*, 1999, pp. 139–142.

[105] J. R. Glass, T. J. Hazen, and I. L. Hetherington, "Real-time telephone-based speech recognition in the Jupiter domain," in *International Conference on Acoustics, Speech and Signal Processing*, 1999, pp. 61–64.

[106] H. J. G. A. Dolfing and I. L. Hetherington, "Incremental language models for speech recognition using finite-state transducers," in *Automatic Speech Recognition Understanding*, 2001, pp. 194–197.

[107] T. J. Hazen, I. L. Hetherington, and A. Park, "FST-based recognition techniques for multilingual and multi-domain spontaneous speech," in *Eurospeech*, 2001, pp. 1591–1594.

[108] G. Zweig, "Speech recognition using dynamic Bayesian networks," Ph.D. dissertation, University of California, Berkeley, 1998.

[109] A. Jansen and P. Niyogi, "Point process models for event-based speech recognition," *Speech Communication*, vol. 51, pp. 1155–1168, December 2009.

[110] L. Rabiner, "A tutorial on HMM and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[111] J. A. Bilmes, "A gentle tutorial on the EM algorithm including gaussian mixtures and baum-welch," International Computer Science Institute, Berkeley, CA, Tech. Rep. TR-97-021, April 1998.

[112] M. Ostendorf, V. V. Digilakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, September 1996.

[113] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[114] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, vol. 4, pp. 127–165, 1990.

[115] J. Allen, *Natural Language Understanding*. Menlo Park, CA: Benjamin and Cummings, 1987.

[116] C.-H. Lee, E. Giachin, L. Rabiner, R. Pieraccini, and A. Rosenberg, "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 6, pp. 103–127, 1992.

[117] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. Mcdonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, no. 2-4, pp. 209–224, 1998.

[118] A. Schmidt-Nielsen and T. H. Crystal, "Human vs. machine speaker identification with telephone speech," in *International Conference on Spoken Language Processing*, 1998.

[119] Y. K. Muthusamy, N. Jain, and R. A. Cole, "Perceptual benchmarks for automatic language identification," in *International Conference on Speech and Signal Processing*, 1994, pp. 333–336.

[120] C.-H. Lee and L. R. Rabiner, "A frame-synchronous network search algorithm for connected word recognition," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 37, no. 11, pp. 1649 – 1658, November 1989.

[121] M. Siafarikas, I. Mporas, T. Ganchev, and N. Fakotakis, "Speech recognition using wavelet packet features," *Journal of Wavelet Theory and Applications*, vol. 2, no. 1, pp. 41–59, 2008.

[122] M. Hasegawa-Johnson, J. Baker, S. Greenberg, K. Kirchhoff, J. Muller, K. Sönmez, S. Borys, K. Chen, A. Juneja, K. Livescu, S. Mohan, E. Coogan, and T. Wang, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop," Johns Hopkins University, Tech. Rep., 2004.

[123] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland College Park, December 2004.

[124] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. 17th International Conference on Machine Learning*, 2000, pp. 591–598.

[125] H. K. J. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 1–6.

[126] J. Bilmes, "Buried Markov models for speech recognition," in *ICASSP*, March 1999, pp. 713–716.

[127] W. Yeyi and A. Acero, "Evaluation of spoken language grammar learning in the ATIS domain," in *ICASSP*, 2002, pp. 41–44.

[128] D. Jurasky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computation Linguistics and Speech Recognition.* Upper Saddle River, NJ: Prentice-Hall, 2000.

[129] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, January 1996.

[130] J. Wu, J. Droppo, L. Deng, and A. Acero, "A noise-robust ASR front-end using Wiener filter constructed from MMSE estimation of clean speech and noise," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 321–325.

[131] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proceedings of ICSLP*, 2002.

[132] D. Dimitriadis, J. C. Segura, L. Garcia, A. Potamianos, P. Maragos, and V. Pitsikalis, "Advanced front-end for robust speech recognition in extremely adverse environments," in *Interspeech*, 2007.

[133] H. Hermansky and N. Morgan, "RASTA processing of speech," in *IEEE Transactions on Speech and Audio Processing*, vol. 2, October 1994, pp. 578–589.

[134] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proceedings of ARPA Speech and Natural Language Workshop*, 1993, pp. 69–74.

[135] S. Furui, "Speaker independent isolated word recognition based on dynamics emphasized cepstrum," *Transactions IECE of Japan*, vol. 69, no. 12, pp. 1310–1317, December 1986.

[136] A. Zolnay, R. Schluter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *International Conference on Acoustics Speech and Signal Processing*, 2005, pp. 457–460.

[137] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[138] J. L. Flanagan, *Speech Analysis and Perception*, 2nd ed. New York, NY: Springer-Verlag, 1972.

[139] H. Hermansky, "Human speech perception: Some lessons from automatic speech recognition," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2001, pp. 187–196.

[140] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *International Conference on Acoustics Speech and Signal Processing*, vol. 1, 1998, pp. 289–292.

[141] S. A. Shamma, X. Yang, and K. Wang, "Auditory representation of acoustic signal," *IEEE Information Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[142] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY: John Wiley and Sons, 2001.

[143] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.1)," Cambridge University Engineering Department, Cambridge, UK, Tech. Rep., 2002.

[144] K. F. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[145] H. Mei-Yuh, X. Huang, and F. A. Alleva, "Predicting unseen triphones with senones," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 6, pp. 412–419, 1996.

[146] L. R. Bahl, P. de Souza, P. Gopalakrishnan, D. Namahoo, and M. Picheny, "Decision trees for phonological rules in continuous speech," in *International Conference on Acoustics Speech and Signal Processing*, vol. 1, 1991, pp. 185–188.

[147] L. R. Welch, "Hidden Markov models and the Baum-Welch algorithm," in *IEEE Information Theory Society Newsletter*, vol. 53, no. 4, Dec 2003.

[148] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[149] X. L. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 16, pp. 89–114, 2002.

[150] R. Kuhn, "Speech recognition and the frequency of recently used words: A modified Markov model for natural language," in *Proceedings of the International Conference on Computational Linguistics*, vol. 1, Budapest, July 1988, pp. 348–350.

[151] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *International Conference on Acoustics Speech and Signal Processing*, vol. 2, 1993, pp. 45–48.

[152] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, pp. 1279–1296, Aug 2000.

[153] B. Bringert, "Speech recognition grammar compilation in grammatical framework," in *Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing*, 2007, pp. 1–8.

[154] P. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *International Conference on Spoken Language Processing*, December 1998, pp. 2711–2714.

[155] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 1043, 1997.

[156] M. Ferras, C.-C. Leung, C. Barras, and J.-L. Gauvain, "Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1366–1378, August 2010.

[157] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, November 2000.

[158] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 1993, pp. 558–561.

[159] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[160] M. Ferras, C. C. Leung, C. Barras, and J.-L. Gauvain, "Constrained MLLR for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 4, Honolulu, April 2007, pp. 53–56.

[161] R. Kuhn, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proceedings of the International Conference for Spoken Language Processing*, vol. 5, 1998, pp. 1771–1774.

[162] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611–622, 1999.

[163] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *Proceedings of Interspeech*, 2005, pp. 217–220.

[164] D. J. Atkinson and A. A. Catellier, "Intelligibility of selected radio systems in the presence of fireground noise: Test plan and results," National Telecommunications and Information Administration, Boulder, Colorado, Tech. Rep. TR-08-453, June 2008.

[165] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.

[166] P. Boersma and D. Weenink, "PRAAT: Doing phonetics by computer," 2009, http://www.praat.org/.

[167] C. F. Hockett, *A Manual of Phonology*. Baltimore, MD: Waverly Press, 1955.

[168] D. Carter, "An information-theoretical analysis of phonetic dictionary access," *Computer Speech and Language*, vol. 2, pp. 1–11, March 1987.

[169] D. Surendran and P. Niyogi, "Quantifying the functional load of phonemic oppositions, distinctive features and suprasegmentals," in *Current Trends in the Theory of Linguistic Change*, O. N. Thomsen, Ed. Amsterdam, Netherlands and Philadelphia, PA: Benjamins, 2009, pp. 43–58.

[170] D. Surendran and P. Niyogi, "The functional load of tone in Mandarin is as high as vowels," in *Speech Prosody*, Nara, Japan, 2004.

[171] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[172] C. E. Shannon, "Prediction and entropy of printed English," *The Bell System Technical Journal*, vol. 30, pp. 50–64, January 1951.

[173] M. Picheny, N. Durlach, and L. Braida, "Speaking clearly for the hard of hearing i. Intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, March 1985.

[174] M. Picheny, N. Durlach, and L. Braida, "Speaking clearly for the hard of hearing ii. Acoustic characteristics of clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 29, pp. 434–446, December 1986.

[175] K. L. Poort, "Stop consonant production: An articulation and acoustic study," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[176] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *International Conference on Acoustics, Speech and Signal Processing*, May 1982, pp. 1291–1294.

[177] S. E. Levinson, "Continuously variable duration hidden Markov models for speech signals," in *International Conference on Acoustics, Speech and Signal Processing*, 1986, pp. 1241–1244.

[178] S. Furui, "On the role of spectral transition for speech perception," *Journal of the Acoustical Society of America*, vol. 80, pp. 1016–1025, October 1986.

[179] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.