Linguistic and Acoustic Analysis of Voice Disguise by Impersonators

Talal Bin Amin

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2015
Acknowledgements

I would like to take this opportunity to acknowledge all the people in my life, near and far, who supported me throughout this life-fulfilling achievement.

I would like to thank my PhD supervisor Assoc. Prof. Pina Marziliano for encouraging me to pursue my research interest in voice impersonation even after my first supervisor Assoc. Prof. Foo Say Wei left the School of EEE. She not only proposed the topic of Linguistic and Acoustic Analysis of Voice Impersonations but she introduced me to members of the Division of Linguistics and Multilingual Studies in the School of Humanities and Social Sciences and established a collaboration with Asst Prof. James Sneed German who later became my co-supervisor. Pina and James both played a vital role in the success and accomplishment of this truly interdisciplinary work. Their advice and knowledgeable comments have always been a source of inspiration and guidance. In the course of my PhD, I have learned many aspects of student-supervisor relationship from them which will be helpful in my future.

I would like to thank the voice impersonators Noella Menon, Marc X Grigoroff and Rishi Budhrani without whom the study would not have been possible.

Also, I would like to acknowledge Charmaine Hon and Zhixin Tan for their efforts in segmentation of the data. I would also like to thank Leow-How Seok Lai, the lab executive, for being very cooperative and helpful in providing the lab equipment.

I also made very good friends during the course of my study. A big thanks to Amrish Nair and Samuel Deslauriers-Gauthier, I will always remember the fun and mostly useful discussions we had over coffee and hope we still keep in touch. I would also like to acknowledge my colleagues Tian Jing, Nikhil Narayan, Nini Soe, Zhang Kunlei, Srinivasan Yadunathan and Lavanya Baskaran for making my experience a memorable one. I will always remember the group outings we had together including the cycling trip in Pulau Ubin and the BBQ dinner for celebrating the completion of my qualifying exam. I am also thankful to Tan Cheen Hau for being a good friend and always open to discussions.

I am grateful to Wong Soo Ting, with out her support and encouragement, I would not have not been able to make it this far.

Finally, I am indebted and grateful to my parents, my brother and my two sisters, their love and support has always been a strong pillar in my life and I owe every success in my life to them.
Abstract

Human voices are distinct and rich with information such as the age, gender, emotional state and identity of the speaker. Voice impersonators possess a great deal of flexibility over their voices which allows them to imitate various people or characters. The question as to how impersonators are able transform their voices and what linguistic and acoustic parameters they rely on is still relatively unexplored. Understanding how they are able to transform their voices holds the key for many applications, such as, speaker recognition, voice transformation, voice disguise detection and speech coding etc.

In the first part of the thesis, the extent to which professional voice artists are able to modulate their voices in order to produce distinct and natural sounding voice identities was investigated. For this purpose, a database of voice impersonations was first constructed using data from three professional voice artists (one male, two females). Each artist produced 9 different voice identities including their natural voice. The data included synchronous speech and electroglottograph signals. The electroglottograph signals provide useful insights into the complex periodic movements of the vocal folds. An acoustic and linguistic analysis was then performed to understand how various glottal parameters such as pitch, vocal fold timing (open quotient through electroglottograph signals), speech rate and vocal tract formants are manipulated by the artists. The analysis revealed that the artists utilized variation in both their glottal and vocal tract characteristics for impersonating different ages and genders. The glottal measures were found to be highly correlated with the perceived age and gender of the impersonated voices. In a novel finding, the artists were found to make changes to their vowel formants on a vowel-by-vowel basis. It was found, in terms of vowel space variability, that the artists were also more consistent with their natural voices as compared to their disguised voice. A listening experiment revealed that the artists were highly successful in deceiving humans which could only correctly identify 56% of the disguised voices. A new objective metric of voice naturalness was proposed which correlates highly with the subjective naturalness ratings of the voices. We also highlight the various constraints involved and the space available to a speaker for producing natural sounding impersonations.

A novel method for the analysis of electroglottograph signals is also introduced. This method models the electroglottograph signal as a sparse signal and allows for the automatic and reliable extraction of the glottal opening and closing instants. Compared to existing methods, this novel method models the glottal opening and closing instants as non-bandlimited signals (diracs) and thus provides
more accurate estimates of their timings.

Voice impersonations also present a challenge for forensic and biometric systems. The final part of the thesis builds upon the linguistic and acoustic analysis and focuses on two biometric applications. The first application aims to automatically discriminate disguised voices of speakers from their natural voices. Acoustic variability related to vowel variances in the F1-F2 space was used as a novel feature for this purpose. This feature was used together with a quadratic discriminant classifier for automatic voice disguise detection. The proposed method was found to outperform the state-of-the-art methods.

For the second application, the goal was to uncover the identity of speakers from both their natural and disguised voices. We proposed a novel method for forensic speaker recognition which uses a phonetic speaker modeling approach for feature extraction and then identifies speakers using the extreme learning machine classifier. This new model requires a very short duration of speech (a frame of 25 ms) for recognition and was found to be more robust than other speaker recognition models. We also investigated and showed how different phonetic units of speech offer different amounts of speaker recognition accuracy.
## Contents

Acknowledgements ........................................... i

Abstract ..................................................... iii

List of Figures ............................................... vii

List of Tables ............................................... ix

Acronyms ..................................................... xi

1 Introduction ............................................... 1

1.1 Contributions ........................................... 3

1.2 Organization of the thesis ................................. 4

2 Background .................................................. 7

2.1 Human Speech Production ................................. 7

2.2 Source Filter Theory of Speech Production .............. 11

2.3 A review on the analysis of Voice Impersonations ....... 13

2.4 A theory of constraints and variability of voices ......... 16

2.5 A review of Voice disguise detection ..................... 18

2.6 Phonetic modeling for speaker recognition ............... 20

3 Glottal and Vocal Tract Analysis of Voice Impersonations 23

3.1 Analysis of a single impersonator .......................... 23

3.1.1 Speech Data ............................................. 24

3.1.2 Results .................................................. 25

3.1.3 Conclusion .............................................. 32

3.2 Analysis of three impersonators ............................ 32

3.2.1 Data Collection .......................................... 32

3.2.2 Analysis and Results ................................... 34

3.2.3 Voice impersonation as a case of voice disguise ..... 52

3.3 Conclusion .................................................. 56

4 Detecting Voice Disguise from Speech Variability .......... 59

4.1 Vowel space analysis ...................................... 59

4.2 Methods .................................................... 61

4.2.1 MFCC-GMM method .................................... 62
5 Forensic Speaker Recognition using the Phonetic GMM-Extreme Learning Machine method 69
5.1 Single GMM based speaker recognition 70
5.2 Phonetic GMM speaker recognition 70
5.2.1 Phonetic GMM-Matched model 71
5.2.2 Phonetic GMM-Unmatched model 71
5.3 Phonetic GMM-Extreme Learning Machine method 74
5.3.1 Extreme Learning Machine classifier 75
5.3.2 The PGMM-ELM model 77
5.3.3 The PGMM-WELM model 79
5.3.4 The PGMM-RELMM model 79
5.4 Experimental setup 80
5.5 Discriminative capabilities of phones 80
5.6 Comparison of the performance of different speaker recognition models 84
5.7 Conclusions 87

6 Glottal Activity Detection using Finite Rate of Innovation Methods 89
6.1 Modeling the Glottal Closing and Opening Instants as Finite Rate of Innovation Signals 91
6.1.1 Recovering the GCI and GOI from a DEGG signal 91
6.2 The proposed scheme 93
6.2.1 Voiced/unvoiced (V/UV) detection 93
6.2.2 Detecting the GCIs 94
6.2.3 Detecting the GOIs 96
6.3 Results and discussion 96
6.4 Conclusion 98

7 Conclusions and Future Work 99
7.1 Conclusions 99
7.2 Future Work 101

A Speech Materials 103

B List of Author's publications 105

Bibliography 107
List of Figures

2.1 Anatomy of the human vocal system [Flanagan, 1972]............. 8
2.2 Simplified view of the human vocal system [Flanagan, 1972]..... 9
2.3 Different views of the larynx........................................ 10
2.4 Voiced and unvoiced signals in time and frequency domains..... 12
2.5 The source filter operation represented in both the time (top row) and frequency domains (bottom row) [Goldstein, February 28, 2012]. ................................................................. 12
2.6 The different constraints within which a speaker can vary his voice. Region 5 shows the space available for variation within which a speaker can produce natural sounding voices.................... 17
3.1 Mean and standard deviation of F0 for the nine voices calculated using Praat............................................................ 26
3.2 Pitch contours for HPF (solid line) and YM (bold line) calculated using Praat............................................................. 27
3.3 The speech rate for the nine voices in terms of syllables per second calculated using [De Jong and Wempe, 2009]. ............... 29
3.4 F1 vs F2 for two different vowels...................................... 30
3.5 The first three formants: F1, F2, F3 for /s:/ in journal for the nine voices................................................................. 31
3.6 The F0 for the the young and old voices of all three impersonators. 36
3.7 The F0 for the the male and female voices of all three impersonators. 37
3.8 The speech rate for the the young and old voices of all three impersonators......................................................... 39
3.9 The speech rate for the the male and female voices of all three impersonators......................................................... 40
3.10 The EGG and DEGG signals corresponding to a voiced speech segment with the labeled GCIs and GOIs. ......................... 42
3.11 The open quotient for the the young and old voices of all three impersonators......................................................... 43
3.12 The open quotient for the the male and female voices of all three impersonators......................................................... 45
3.13 The vowel space for the nine impersonated voices of impersonator 1F. The vowel ellipses represent the 95% confidence region... 47
3.14 The vowel ellipses for the target vowels for some voices of impersonator 1F. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

3.15 The vowel ellipses for the target vowels for some voices of impersonator 2M. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

3.16 The vowel ellipses for the target vowels for some voices of impersonator 3M. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

4.1 The natural and disguised voices for vowel /u/. The ellipses represent the 95% confidence region.

4.2 The 95% confidence ellipse for vowel /ae/. The major and minor axis of the ellipse are represented by $\gamma_1$ and $\gamma_2$.

5.1 Flow diagram of the Phonetic GMM-Matched model. The input frame is only compared to its corresponding phone model.

5.2 Flow diagram of the Phonetic GMM-Unmatched model. The input frame is compared to all phone model.

5.3 The architecture of a single-hidden layer feed forward neural network.

5.4 The architecture of the WELM speaker recognition model.

5.5 Receiver Operating Characteristics plot for the detection of natural voices.

5.6 Receiver Operating Characteristics plot for the detection of disguised voices.

6.1 A period of aDEGG signal and the FRI signals $x_1(t)$ and $x_2(t)$. The signals $x_1(t)$ and $x_2(t)$ only capture the points of interest, i.e. the GCI and GOI.

6.2 The flow chart for the GCI and GOI detection scheme using the FRI model of DEGG signals.

6.3 Voiced/unvoiced detection for a DEGG signal corresponding to a sentence. The values of ‘V’ and ‘UV’ on the vertical right hand axis indicate if a DEGG segment is voiced or unvoiced respectively.
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Abbreviations of the nine different voices from the voice artist.</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Mean and standard deviation of F0 in Hertz. The natural voices are shown in bold.</td>
<td>35</td>
</tr>
<tr>
<td>3.3</td>
<td>Mean and standard deviation of the speech rate (syllables per second). The natural voices are shown in bold.</td>
<td>38</td>
</tr>
<tr>
<td>3.4</td>
<td>Mean and standard deviation of open quotient. The natural voices are shown in bold.</td>
<td>44</td>
</tr>
<tr>
<td>3.5</td>
<td>Subjective and Objective ratings of the voices of the three impersonators.</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>The confusions matrices for the different matrices indicating their performance for disguised and natural voices. A high performing system has higher values along the diagonal entries and lower values of the off diagonal entries.</td>
<td>66</td>
</tr>
<tr>
<td>5.1</td>
<td>The phones used during analysis and their time duration.</td>
<td>81</td>
</tr>
<tr>
<td>5.2</td>
<td>The recognition rates (RR) of the phones in percentage. The results are shown here for both the natural and disguised voices of impersonators obtained from the PGMM-M and PGMM-U models.</td>
<td>82</td>
</tr>
<tr>
<td>5.3</td>
<td>Testing accuracy in percentage for the natural and disguised voices.</td>
<td>85</td>
</tr>
<tr>
<td>5.4</td>
<td>Area under curve for the natural and disguised voices.</td>
<td>87</td>
</tr>
<tr>
<td>6.1</td>
<td>GCI detection performance on the APLAWD database</td>
<td>97</td>
</tr>
<tr>
<td>6.2</td>
<td>GOI detection performance on the APLAWD database</td>
<td>97</td>
</tr>
</tbody>
</table>
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>DEGG</td>
<td>Differentiated Electroglottograph</td>
</tr>
<tr>
<td>EGG</td>
<td>Electroglottograph</td>
</tr>
<tr>
<td>ELM</td>
<td>Extreme Learning Machine</td>
</tr>
<tr>
<td>FRI</td>
<td>Finite Rate of Innovation</td>
</tr>
<tr>
<td>GCI</td>
<td>Glottal Closing Instant</td>
</tr>
<tr>
<td>GOI</td>
<td>Glottal Opening Instant</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate Analysis of Variance</td>
</tr>
<tr>
<td>MFCCs</td>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>PGMM</td>
<td>Phonetic Gaussian Mixture Model</td>
</tr>
<tr>
<td>PGMM-M</td>
<td>Phonetic Gaussian Mixture Model-Matched</td>
</tr>
<tr>
<td>PGMM-U</td>
<td>Phonetic Gaussian Mixture Model-Unmatched</td>
</tr>
<tr>
<td>PGMM-ELM</td>
<td>Phonetic Gaussian Mixture Model-Extreme Learning Machine</td>
</tr>
<tr>
<td>PGMM-WELM</td>
<td>Phonetic Gaussian Mixture Model-Weighted Extreme Learning Machine</td>
</tr>
<tr>
<td>PGMM-RELM</td>
<td>Phonetic Gaussian Mixture Model-Reduced Extreme Learning Machine</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristics</td>
</tr>
<tr>
<td>SGMM</td>
<td>Single Gaussian Mixture Model</td>
</tr>
</tbody>
</table>
Introduction

Voice impersonation is an art which involves changing one’s voice to sound like another person. It is mostly done for entertainment purposes, e.g. caricaturization and in media related fields. However, the study of voice impersonators is also important in other fields of research including forensics [Hollien, 2002], voice disguise detection [Perrot et al., 2007], speaker recognition [Bonastre et al., 2003], text-to-speech (TTS) synthesis and voice conversion [Stylianou, 2009] etc. From the point of view of forensics, for example, one can mask one’s identity through voice disguise in order to avoid being identified. It is therefore important to develop methods that allow law enforcement authorities to identify individuals in spite of such modifications. Similar issues apply to security systems based on speaker recognition, which are also vulnerable to circumvention by voice impersonators [Zetterholm, 2006, Farrús et al., 2010]. Currently, many speech transformation applications such as voice conversion [Machado and Queiroz, 2010, Stylianou, 2009] and Text-To-Speech (TTS) synthesis also suffer from a lack of naturalness in the synthesized speech. This suggests that key aspects of speech that contribute to naturalness are being ignored by current speech transformation techniques. There exists a need, therefore, to identify the set of voice parameters that are involved in successful (i.e., natural-sounding) voice modification, and to explore how these can be used to improve voice disguise identification, voice identity detection, as well as how these parameters can be manipulated in a way that preserves naturalness across transformations.

A direct inspiration can be drawn from voice impersonators, who successfully maintain both naturalness and individuality while producing different voice identities. Such impersonations can be natural enough to deceive humans as well as automated speaker identification systems [Zetterholm, 2006, Farrús et al., 2010].

To understand why voice impersonators are successful and what parameters they rely on to change their voice, we first collected a database consisting of
synchronous speech and ElectroGlottoGraph (EGG) signals obtained from three voice impersonators each producing nine distinct voice identities. The aim was to elicit a large degree of variation approximating the maximal parameter space available to the speakers. We analyzed both glottal and vocal tract measures including pitch (F0), speech rate, the vowel formant frequencies (F1, F2), and the timing characteristics of the vocal folds.

In the first part of this thesis, we seek to determine which vocal and acoustic parameters the impersonators make use of in order to achieve different voice identities, and to a certain extent, the relationship of those parameters to specific identity traits indicated by the associated labels (e.g., age or gender). The impersonators in this study impersonated various fictitious characters rather than specific target speakers, thereby allowing to investigate the issues relating to voice disguise rather than similarity. The parameters we chose to investigate are in fact largely motivated by what is already believed to have implications for voice identity (e.g., F0 or pitch is indicative of gender since female speakers are generally associated with a higher mean F0 and greater temporal variation than males [Hillenbrand and Clark, 2009, Honorof and Whalen, 2010, Sambur, 1975]), though identifying such associations was not the central goal of this study since these are mostly well-known. Instead, we seek to explore the issue of how large the space of variation is within the constraints of naturalness, and how this is influenced by both speaker-specific traits as well as linguistic structure (in this case, the structure of the English vowel inventory).

We also report the results of a subjective test by naive listeners that provides an estimate of how well the impersonators were able to disguise their voices and relates this to their natural voice productions. Since our findings on vowel formants reveal an important effect of vowel category dependency, we introduce a no-reference objective measure for voice disguise that accounts for such effects. We found that acoustic variability, such as that, associated with vowel variability in the formant space contrasted between disguised and natural voices of speakers. The resulting scores of this objective metric are then compared against those of the listening test.

For the analysis of vocal cords movements, we introduce a novel algorithm which accurately determines the glottal closing instants (GCIs) and glottal opening instants (GOIs) from the Differentiated ElectroGlottoGraph (DEGG) signals. This method models the DEGG signal as a non-bandlimited signal having two Diracs in each period which correspond to the GCI and GOI. Methods for sampling and reconstructing signals with Finite Rate of Innovation (FRI) are then used to recover the GCIs and GOIs from the DEGG signal.

A new theory of constraints and variability of voices is also introduced. This theory illustrates how the various factors: physical, linguistic and naturalness offer different spaces of variation and how a speaker is bounded by the constraints imposed by them in order to produce natural sounding impersonations.
1.1. CONTRIBUTIONS

Building upon the analysis, we also present two important applications in the context of forensics and biometrics: a) automatic detection of voice disguise and b) forensic speaker recognition, that is, recognizing speakers from their disguised voices. The field of automatic speaker recognition has seen significant improvements in recent years, though it still suffers from several limitations. In the context of forensics, speaker recognition is still done manually by phoneticians using aural-spectrographic inspection, a method which is both highly labor intensive and subject to bias. This is largely due to the fact that automated systems still do not adequately address the problem of voice disguise [Bonastre et al., 2003, Campbell et al., 2009]. In fact, most forensic cases involve a criminal disguising his or her identity through voice disguise. A disguise identification system can therefore provide a front end to a speaker recognition system by giving an indication that the voice is disguised before actually attempting to determine the speaker’s identity. Such a system has the potential to reduce computational resources as well as to facilitate the automation of speaker recognition systems more generally. All current methods of automated voice disguise detection [Perrot et al., 2007, Mary et al., 2012] that we are aware of make use machine learning methods and require large amounts of training data to achieve reasonable performance. However, in real forensic scenarios the amount of available data is usually very limited. Thus there exists a need for an objective metric that evaluates voice disguise without requiring any training data or a reference.

In a second application, we explore ways to identify speakers from their disguised voices. Thus, once a voice has been identified as being disguised by a disguise detection system, the next step is to identify the true identity of the speaker. Since, in practice, it is not possible to have details about how a person will disguise his/her voice, our goal here is to find methods which only use the natural voice samples for training and then can predict the identity of the speaker from the disguised voice samples. Here, we make use of the phonetic approaches for speaker modeling. We also explore how different phones categories contribute differently to the accuracy of a speaker recognition system.

1.1 Contributions

The contributions of this thesis are as follows:

- Building a database of voice impersonations and analyzing the flexibility of speakers in terms of formant frequencies, speech rate, vocal cord movements using electrogolottograph signals and F0. Also, revealing the specific glottal and vocal tract features the artists rely on to achieve distinct voice identities and that proposing a novel objective metric of voice naturalness related to vowel variances which correlates with the subjective rating of naturalness of voices.
• Proposing a new theory of constraints and variability which shows how the physical, linguistic and naturalness factors influence the space available to a speaker for impersonating and also highlights the constraints imposed by them.

• Proposing a new quantitative feature related to vowel variances which together with a linear discriminant classifier can be used for automatically discriminating disguised and natural voices of speakers. This novel method outperforms traditional spectral based methods.

• Proposing a novel forensic speaker recognition model which utilizes the segment level phonological structure for identifying speakers from both their natural and disguised voices. The model uses only the natural voices of speakers for training and makes no assumptions about the kinds of voice disguise. Several variations of the proposed model are also introduced which improve the recognition performance and computational efficiency.

• Proposing a novel sparse model for electroglottograph signals and developing a new scheme for automatic and reliable extraction of glottal closing and glottal opening instants.

1.2 Organization of the thesis

The remainder of the thesis is organized as follows: Chapter 2 provides a brief account of the speech production mechanism from both a physiological and signal processing perspective. We then explain the importance of performing a linguistic and acoustic analysis of impersonations and its usefulness for various applications. We also discuss how acoustic variability, specifically, vowel variances, is an important factor in many different areas of speech analysis and why it can be useful for forensic applications, such as, disguise detection. Finally, we also show how phonetic modeling is useful for understanding speakers ability in modifying their voices and why it can be important for forensic speaker recognition.

Chapter 3 provides a linguistic and acoustic analysis of voice impersonations. First, the impersonations from a single artist are analyzed. The parameters which are studied include pitch, speech rate and vowel formants. To better generalize our findings, the analysis is then extended to include impersonations from three different voice artists. We also study the vocal cord movements through electroglottograph signals. Next, the results from a subjective listening experiment, where subjects rated the naturalness of the impersonations, are presented. An objective metric related to vowel variances in the formant space is introduced. The objective metric is compared to the subjective ratings of naturalness.

Chapter 4 introduces a method for automatically discriminating between disguised and natural voices. We build upon the ideas presented in Chapter 3 and
use the acoustic variability related to vowel variances in the formant space as a robust feature to distinguish between disguised and natural voices of speakers. The feature is then used together with the linear discriminant analysis classifier for disguise detection and compared against the state-of-the-art methods.

The phonetic modeling of speakers is presented in Chapter 5. The goal here is to be able to identify speakers from their disguised voices while only having access to their natural voices in training. Several different models which rely on phonetic modeling of speakers are introduced and compared against existing speaker recognition models. A ranking of phones in terms of their speaker discriminating performance is also presented.

In Chapter 6, a novel and robust method for the detection of glottal opening and closing instants from electroglottograph signals is presented. This method was used for analyzing the vocal cord movements as discussed in Chapter 3. The proposed method uses the finite rate of innovation methods to model the glottal opening and closing instants as non-bandlimited signals, specifically as diracs. The proposed algorithm is then tested against various existing methods for performance evaluation.

Finally, Chapter 7 concludes the thesis and possible extensions of the work are discussed.
Background

This chapter first provides a brief account of the human speech production system. In Section 2.1, a simple explanation of the human speech production mechanism from a physiological point of view is presented. This is followed by a signal processing model of human speech production described in Section 2.2. This chapter also provides a literature review of the various concepts discussed in this study together with the motivations of the different approaches undertaken. First, a review on the analysis of voice impersonations is presented in Section 2.3. Next, a theory of constraints and variability of voices is presented in Section 2.4. This is followed by a review of voice disguise detection and phonetic categories approaches for speaker recognition given in Section 2.5 and Section 2.6 respectively.

2.1 Human Speech Production

The process of human speech production is well understood from a physiological point of view, however, to model it accurately in mathematical terms is still an open problem. In this section, a simplified version of the human speech production mechanism is presented. The interested reader can refer to [Stevens, 2000] for a more detailed explanation.

The human speech production system can be divided into three basic subsystems which are the subglottal area, the larynx and the supraglottal area. Figure 2.1 shows the cross-sectional view of the human vocal system. The subglottal area is below the larynx and consists of the diaphragm, the lungs and the trachea. The supraglottal area is located above the larynx and comprises the larynx tube, the pharyngeal cavity and the oral and nasal cavities. The supraglottal area is also
known as the vocal tract. The tongue, jaws, lips and velum are the articulators and are used for the production of different speech sounds.

A simplified version of the speech production system is shown in Figure 2.2. The lungs act as a power source and generate the airflow which travels through the trachea. This airflow then reaches the larynx. The larynx is the organ which acts as the source of sound and consists of the vocal folds also commonly known as the vocal cords. Figure 2.3 shows two different views of the larynx. The space between the vocal folds is known as the glottis or the voice box.

For the production of voiced sounds, the airflow from the lungs gets modulated in a quasi-periodic manner by the vocal folds. A widely accepted theory of vocal fold vibration is known as the myoelastic-aerodynamic theory of voice production [Berg, 1958] which is described here. The vocal folds act as a constriction in the path from the lungs to the vocal tract. At first, the vocal folds are in a closed state, once the air pressure below the glottis builds to a sufficient level the vocal folds open thereby resulting in the air flowing through the glottis. According to Bernoulli’s principle, when a fluid flows through an orifice, the pressure is lower in the constriction than on either side. The pressure in the glottis falls due to Bernoulli’s principle resulting in closing of the vocal folds. The resultant quasi-periodic air flow then passes through the vocal tract. The same cycle repeats again. The resulting waveform after passing through the glottis is known as the glottal waveform. The glottal waveform represents the changes of air pressure
Figure 2.2: Simplified view of the human vocal system [Flanagan, 1972].
2.1. HUMAN SPEECH PRODUCTION

(a) Transversal view of the larynx [Gray, 1918].

(b) Front vertical cut of the larynx [Degottex, 2010].

Figure 2.3: Different views of the larynx.
2.2. SOURCE FILTER THEORY OF SPEECH PRODUCTION

over time. The glottal waveform is also known as glottal source or excitation. This glottal waveform then passes through the vocal tract which is the supraglottal area shown in Figure 2.1 and comprises of the larynx tube, pharyngeal cavity, articulators (tongue, lips, teeth, jaw etc.), and the oral and nasal cavities. The vocal tract changes shape continuously to produce different sounds. The particular shape of the vocal tract at any time depends primarily on the sound the speaker is trying to produce, and different shapes give rise to different resonant frequencies called formants. The vocal tract contains two openings: the nasal and the oral cavity. The velum position as shown in Figure 2.2 controls the amount of nasality. When the velum is lowered the sound exits through the nasal cavity and vice versa.

For unvoiced sounds, the glottis is kept open and the airflow passes freely through the larynx without any obstacles. However, a constriction is created at some point in the vocal tract. This constriction creates turbulence and gives rise to a noise-like excitation. The sound created via a constriction is described as a noise source. It contains no dominating periodic component and has a relatively flat spectrum. Periodic and aperiodic sources can be generated simultaneously to produce mixed voiced and aperiodic speech typical of sounds, such as, voiced fricatives. The waveforms for voiced and unvoiced speech signals are shown in Figure 2.4. It can be seen that the vowel /oʊ/ is periodic as shown in Figure 2.4(a) and Figure 2.4(b) respectively. However, the unvoiced sound as shown in Figure 2.4(c), that is, the fricative /f/ has noise like characteristics. It has a power spectrum which is relatively flat meaning that the power is distributed over all the frequencies as can be seen in Figure 2.4(d).

2.2 Source Filter Theory of Speech Production

In order to model the human speech production using a signal processing approach, the glottal waveform and the vocal tract are assumed to be independent of each other. Although, in reality they interact in a nonlinear fashion [Titze, 2008]. The relationship between the two, however, is not well understood mathematically. Therefore, for simplicity, most signal processing based methods assume that they are independent. According to the acoustic theory of speech production [Fant, 1970], the human speech production process can be viewed as a source and filtering operation. The larynx acts as the source and the vocal tract acts as a time varying filter. This can be represented mathematically as

\[ s(t) = \int_{0}^{t} h(t, t - \tau)u(\tau) \, d\tau \]  

\[ (2.1) \]

where \( s(t) \) is the speech signal, \( h(t, t - \tau) \) is the time-varying vocal tract impulse response and \( u(\tau) \) is the excitation or the source signal.

For the production of voiced sounds, a series of periodic impulses are pro-
2.2. SOURCE FILTER THEORY OF SPEECH PRODUCTION

Figure 2.4: Voiced and unvoiced signals in time and frequency domains.

(a) Vowel /ou/ in time domain.  
(b) Power Spectrum of Vowel /ou/.

(c) Fricative /f/ in time domain.  
(d) Power Spectrum of Fricative /f/.

Figure 2.5: The source filter operation represented in both the time (top row) and frequency domains (bottom row) [Goldstein, February 28, 2012].
duced at the source. The vocal cords vibrate in a quasi-periodic manner. The source in voiced speech results from the vibration of the vocal folds in response to the airflow from the lungs. The frequency of these pulses is called F0 which is an objective measurement that correlates closely with the perceived pitch of speech. The filter has certain resonant frequencies depending on the shape of the vocal tract. These are called the formant frequencies. The peaks in the vocal tract filter response as shown in Figure 2.5 represent the formant frequencies. The formant frequencies are generally represented as F1, F2, ..., Fn. Here F1 indicates the first formant frequency, F2 indicates the second formant frequency and so on. The amplitude of the harmonics decreases at 6 dB/octave. The harmonics of the source signal which have a frequency close to a resonance frequency of the vocal tract are able to pass freely through the vocal tract and thus produce a formant. However, the harmonics whose frequencies are not close to any vocal tract resonance frequency become weakened and appear as troughs between the formant peaks. Figure 2.5 shows the the source filtering operation in both the time and frequency domains.

Since both the source and filter are responsible for the production of speech, a speaker can modulate various source and filter parameters to produce different voices. Thus, in order to understand how voice impersonators are able to transform their voice identity, it is preferable to analyze both the glottal (source) and vocal tract (filter) characteristics of voice impersonators. The glottal features include F0, timing characteristics of the vocal folds through EGG signals, speech rate. The vocal tract features studied are the vowel formants. Next, an overview of the analysis on voice impersonations, motivations and background is presented.

2.3 A review on the analysis of Voice Impersonations

Voice impersonators possess a flexible voice which they can use to change their voice identity. This flexibility allows them to imitate people and characters that differ in age, gender, regional accent or voice quality. These impersonations are highly natural and readily distinguished from the impersonator’s original voice. This presents a challenge for forensic analysis and speaker identification techniques, which are prone to attacks of voice disguise. The phenomena underlying a successful voice impersonation, however, are not well understood. Little is known, for example, about the inventory of glottal and vocal tract features that impersonators are able to exploit, and for features that are already recognized, the amount and type of variation that is possible for any one speaker given constraints of anatomy (e.g., vocal tract length), naturalness and linguistic structure (e.g., vowel space organization).

The analysis of the glottal and vocal tract parameters of impersonated voices
can be useful for voice disguise identification, where there is a need to identify a set of parameters that can be helpful in determining whether a voice is disguised or not. In this regard, the analysis and comparison of an impersonator’s natural voice with the impersonated (disguised) voices reveals how various acoustic parameters are manipulated to extend a space of disguised voices around their natural voice. It can also reveal any invariant parameters or systematic relationships between the natural and impersonated voices, either of which may be readily exploited for biometric application, such as, voice disguise detection and speaker recognition etc. In [Stylianou, 2009] the need for studying voice impersonators was specifically highlighted in connection with voice conversion applications in order to better understand how the issue of naturalness under identity changes can be better incorporated into speech transformation algorithms.

Not all portions of the range of variation that a speaker is capable of producing will result in natural-sounding speech. At the same time, there are limitations on the range of variation that a given speaker can produce. Speech parameters such as F0 (pitch) range and formant frequency, for example, may be constrained by a speaker’s physical traits (esp. vocal cord anatomy and vocal tract length). As a first step, then, it is important to be able to model not only the amount and type of variation within the total parameter space that results in natural-sounding voices, but also to consider which regions of that space are achievable by a single speaker given his or her inherent physical limitations. In that sense, a central goal of our study is to begin to “map out” the space of variation in speech parameters that corresponds to natural-sounding speech, and to do so in a way that takes into account speaker-specific limitations.

Some voice parameters can be important both for speaker identity as well as for the actual linguistic content of the utterances involved. Vowel formant frequencies, for example, are influenced by vocal tract length, and therefore serve as an important cue to a speaker’s age and gender. Simultaneously, it is the relative differences between vowel formant frequencies that ultimately creates the distinction between different vowel sounds (e.g., the difference between the vowels in the words “bed” and “bad”). Crucially, this suggests that the space of variation cannot be correctly modeled without taking into account linguistic structure. A second goal of our study, therefore, is to explore the extent to which the variation exhibited across different natural-sounding voice identities depends on linguistic structure in a systematic way. In essence, we seek to test whether the shape of this parameter “space” is influenced or constrained by specific features of the language involved. More generally, we hope to bring to light previously undocumented challenges faced by current approaches to voice disguise, speaker identification, voice conversion, and speech synthesis, and to identify potential solutions to those challenges. Studies on voice impersonation are limited [Zetterholm, 2006, Eriksen and Wretling, 1997, Kitamura, 2008, Zetterholm, 2009]. The focus of existing studies has been to determine how closely an impersonator can approximate
2.3. A REVIEW ON THE ANALYSIS OF VOICE IMPERSONATIONS

a target speaker, as well as whether the glottal and vocal tract measures exhibit a close correspondence. Different data sets have led to different observations in this regard. For example, in [Eriksson and Wretling, 1997], 30-second excerpts of uninterrupted Swedish sentences were analyzed, while in [Kitamura, 2008], only two short Japanese sentences were used. In [Zetterholm, 2006], different sentences were used for different target voices, and only one word was common to all sentences and therefore useful for comparison. Additionally, the sentences used in [Zetterholm, 2006] were designed to be humorous and therefore lacked emotional neutrality, a fact which may have confounded or masked the effects of voice identity. In [Eriksson and Wretling, 1997], it was concluded that the voice impersonator found it difficult to accurately modify vocal tract characteristics towards the target speaker, whereas in [Kitamura, 2008] the impersonator was able to modify both the prosodic and vocal tract characteristics towards the target speaker. The different outcomes among these studies could be attributed to the fact that the impersonators had different skill sets, and different voice targets to imitate in different languages. The findings from such studies are also hard to be generalizable without having a larger sample size and more number of impersonators.

In previous studies, the goal of the impersonator was to imitate the voices of specific speakers. This is in contrast to our study, where the impersonators creatively adapted their voices to produce character voices from their own repertoire. While the impersonators gave labels to some of these voices that were indicative of certain identity traits (e.g., “high pitch female”), they were not given instructions to target specific identities, voices, or identity traits. This allowed the impersonators to more fully express the flexibility of their voices, by impersonating a wide range of voice identities that they were comfortable producing. This in turn allowed us to explore the issues of variation and naturalness rather than similarity to a target speaker.

Crucially, none of the previous studies have investigated how vocal fold behavior changes when an impersonator produces different voices. For languages like English, vocal fold behavior (e.g., creakiness or breathiness) is largely unimportant for word or sentence meaning, though it is known to be associated to the social identity, especially gender [Wolk et al., 2012]. We therefore hypothesize that our voice impersonators will be able to recruit variation in vocal fold parameters in their attempt to create distinct voice identities. If they cannot, then there is evidence for one or more speaker-specific stable parameters that may be useful for speaker identification or voice disguise identification. Here, we make use of the Electroglottograph (EGG) signal, which provides a direct representation of the vocal fold vibration patterns and is free from the filtering effects of the vocal tract. The EGG signal has been found to be independent of vowel category [Epstein, 2002] and to depend primarily on the anatomical characteristics of a speaker’s vocal folds [Campbell et al., 2003]. While a few studies have shown that the EGG
signal can be used for speaker identification [Campbell et al., 2003], this is, to our knowledge, the first study to use EGG signals for the analysis of voice impersonations.

2.4 A theory of constraints and variability of voices

Given that an impersonator will use variability in his/her voice to produce different impersonations, it is important to understand what factors influence variability and what are the constraints imposed by them. As already highlighted in the previous section, a speaker is limited in his ability to vary his voice by a) physical constraints, b) constraints imposed by the language and c) naturalness constraints. Physical constraints are imposed by the speech production apparatus of a speaker. Linguistic constraints on the other hand are imposed by the language as well as dialect used by the speaker. The linguistic constraints come from a specific community of people (of which the speaker is a part), for example, from those who speak a certain dialect of a language. Linguistic constraints thus include dialectal constraints, pronunciation rules etc. Lastly naturalness constraints are necessary to ensure that a speaker’s voice sounds natural. Here naturalness means that a speaker is not straining his voice beyond the range used in day-to-day communicative speech. The naturalness constraints again come from an individual’s experience with a community of other speakers. The naturalness constraint is extremely important as it is possible for a speaker to satisfy both physical and linguistic constraints and still sound unnatural, for example, when screaming some one’s name.

Each factor: physical, linguistic and naturalness offers a space within which a speaker can vary his speech. The three different spaces of variation are represented by the circles in Figure 2.6. Each circle represents a source of variation and the boundary of a circle represents the constraints imposed by it. There are several regions numbered in Figure 2.6. Region 1 is the space of physical variation which is not constrained by the language. Thus region 1 includes non-speech sounds a speaker can produce. Region 2 is the space of physical variation which satisfies the constraints that allow it to be recognized as meaningful speech, but not natural. A speaker thus produces speech in this region which is unnatural sounding (for example, a male speaking in an extremely high pitch). Region 3 is the space of linguistic variation which is outside the physical and naturalness constraints. Region 3 can thus include, for example, synthetic speech which does not sound like it came from a human and which a human cannot accurately reproduce. The linguistic variation space comes from a community rather than an individual and the speaker only occupies a subset of that space. Naturalness constraints, on the other hand, are a subset of linguistic constraints. Region 4 is the space of naturalness which is only constrained by the language whereas in region 5 it is constrained by both the language and the physiology of the speaker. Region
Figure 2.6: The different constraints within which a speaker can vary his voice. Region 5 shows the space available for variation within which a speaker can produce natural sounding voices.
is thus the space within which a speaker can vary his voice to produce different
natural sounding voice identities. In conclusion, a speaker (impersonator) must
satisfy physical, linguistic and naturalness constraints to successfully transform
his voice.

2.5 A review of Voice disguise detection

The automatic detection of disguised voices is an interesting and challenging
problem. Current speaker recognition systems are becoming increasingly more
accurate. However, their robustness to attacks of voice disguise is still a relatively
unexplored area. Several studies, including [Künzel et al., 2004, Zhang and Tan,
2008], have reported that the presence of disguised voices can significantly re-
duce the performance of speaker recognition systems. For example, when making
a phone call, criminals will typically try to conceal their identity by disguising
their voices. In such cases, having information about the nature of the voice (dis-
guised/ natural) can prove useful for uncovering the true identity of the speaker.
More generally, a disguise detection system can also be utilized for reducing the
misclassification rate of speaker recognition system. By using a disguise detec-
tion system as a front end to a speaker recognition system, disguised voices can
be filtered out before any attempt is made to relate them to the voices of autho-
ized individuals in the system. All current methods of automated voice disguise
detection [Perrot et al., 2007, Mary et al., 2012] that we are aware of use machine
learning methods and require large amounts of training data to achieve reasonable
process. However, in real forensic scenarios the amount of available data is usu-
ally very limited. Thus there exists a need for an objective metric that evaluates
voice disguise without requiring any training data or a reference.

The simplest approach for disguise detection, as was used in [Perrot and Chol-
lett, 2008], is the same as that used for speaker verification [Reynolds et al., 2000].
In the case of speaker verification, the goal is to verify whether a given set of
speech samples belong to the claimed speaker or not. The baseline method [Reynolds
et al., 2000] utilizes Mel Frequency Cepstral Coefficients (MFCCs) as features
which are parameterized representations of the spectral envelope. These are ex-
tracted from short-time frames (25 ms) over the voiced segments of speech sam-
pies. During the training phase, Gaussian Mixture Models (GMM) are used to
build two statistical speaker models. The first model, called the speaker model,
corresponds to the claimed speaker, while the background model theoretically
represents all possible speakers except the claimed speaker. When an unknown
speech sample is presented to the system during testing, the likelihood ratio test
is used for making a binary decision, that is, whether the given speech samples
belong to the claimed speaker model or to the background model. This MFCC-
GMM method works best if the spectral features across the speaker and back-
ground models are well-separated; in other words, the speakers in the two models
have very different spectral characteristics [Reynolds and Rose, 1995]. For the
case of disguise detection, the two models that are built are the disguised and nat-
ural voice models [Perrot and Chollet, 2008]. Although the formulation of the
problem of disguise detection is similar to speaker verification, there are reasons
to believe that using spectral features for disguise detection might not be the best
possible approach. For speaker verification, the assumption is that the spectral
features of the claimed speaker and the rest of the speakers are highly distinct.
However, this assumption might not hold for disguise detection where the spec-
tral features related to the natural and disguised voices of the “same” speaker can
have significant overlap. In particular, some speech segments across the natural
and disguised voices of a speaker may be very similar (poorly-disguised) while
some might be highly dissimilar (well-disguised)

An important aspect of this work is to study the acoustic variability (related
to the vowel space) exhibited by voice impersonators. Any specific differences
in the variability of speakers for their natural and disguised voices can be useful
for disguise detection. Previously, vowel space variability has also been found to
be a relevant factor for speech recognition and speech intelligibility [Wade et al.,
2007]. This is the first work we know of which studies the variability exhibited by
a speaker when impersonating. The closest work to this was in [Weirich, 2010]
where variability of monozygotic (genetically identical) twins was studied. Since
monozygotic twins can be considered to have identical production apparatuses,
social factors, such as learning environment, were attributed to the inter-speaker
variability of twins in the vowel space.

In this study, we expect speakers to demonstrate different amounts and kinds
of variability for their natural and disguised voices, specifically in the vowel space.
Consider, for example, when an American English speaking impersonator tries to
imitate a character voice in a British accent. We can expect the impersonator to
have more familiarity with American English, having had a more rich and natu-
ral learning experience in it compared to British English. The impersonator can
thus produce vowel tokens more reliably (less variability) for American English
compared to tokens when impersonating a character in British English. Without
saying more, this suggests less variability for the native dialect compared to the
non-native dialect. Thus, social factors such as learning experience can play an
important role in the variability exhibited by a speaker when impersonating.

The goal in this study is to use the the variability exhibited by impersonators
to discriminate between their natural and disguised voices. We aim to propose
a new method which addresses the shortcomings of the traditional approaches a)
by modeling the variability of speakers’ productions rather than comparing spec-
tral features directly and b) by focusing only on certain speech segments (vowels
in this case) rather than using all the voiced segments of speech. Some studies,
including [Sturim et al., 2002], have already shown that considering only certain
speech segments for analysis rather than all voiced segments of speech helps to
bring out fine-grained phonetic differences of a speaker’s productions. In other words, we take higher level linguistic information, the vowel category, into consideration as opposed to using only a simple voiced/unvoiced discrimination. Another key advantage of this approach is that less data is needed for modeling the disguised and natural voices. This is particularly useful in forensic scenarios where the amount of data is limited.

2.6 Phonetic modeling for speaker recognition

Speaker recognition is a biometric system which uses speech samples to retrieve the identity of a speaker. Research on improving the performance of speaker recognition systems is still ongoing. Much of the research in speaker recognition has been conducted to improve system performance, for example, when there is a mismatch between the training and testing conditions [Kinnunen and Li, 2010]. These include scenarios, such as, different noise levels in the testing or training conditions, handset mismatch, channel distortions etc. The current speaker recognition systems achieve good performance in similar training and testing conditions [Kinnunen and Li, 2010]. More recently, a few studies, such as [Künzel et al., 2004, Zhang and Tan, 2008], have reported that impersonated (disguised) voices also present challenges to current speaker recognition systems and can significantly reduce their performance. However, no particular study has explored ways to make speaker recognition system robust to attacks of voice disguise.

The baseline method for speaker recognition [Reynolds and Rose, 1995] models each speaker with a GMM using MFCCs as features. This approach does not take any phonetic information into consideration and therefore does not require a phonetic segmentation of the data for training the speaker models. However, as we will demonstrate in Chapter 3, a lot of speaker specific information is contained at a phone (vowel) level. Specifically, we demonstrate how different vowels are subject to different kinds of variations in the vowel space and how the impersonators made adjustments to vowel formants on a vowel by vowel basis. Also, several studies have already reported that different phonemes exhibit different speaker separating capabilities. In [Eatock and Mason, 1994, Faltlhauser and Ruske, 2001], for example, it was found that nasals /n/ and /m/ provided higher identification rates while having relatively lower frequency of occurrence, thereby indicating their superiority over other phonemes. Although, other sophisticated methods for speaker recognition, such as, the Joint Factor Analysis (JFA) model [Kenny et al., 2007, Dehak et al., 2011] exist, the focus of this thesis is on developing phonetic categories approaches for speaker recognition. More recently the I-vector approaches, such as [Larcher et al., 2012], and the phonetically aware Deep Neural Networks [Lei et al., 2014] have also been utilized for speaker verification and recognition tasks. The recent trend in speaker modeling has thus been to incorporate phonetic information to improve system performance.
2.6. PHONETIC MODELING FOR SPEAKER RECOGNITION

Inspired from this, this thesis focuses on developing speaker recognition methods which model different units of speech (phones) separately, so-called phonetic categories approaches for speaker recognition. We expect speakers to disguise some phonemes more easily than others. Nasals, for example, have been found to be both stable and speaker specific as the volume or shape of the nasal cavity is relatively fixed [Amino et al., 2006] and thus harder for speakers to manipulate. On the other hand, vowels, such as /u/, are more variable as the speaker can use variations in lip rounding during production. Thus we expect different phonemes to offer different amounts of speaker recognition accuracy. The phonetic categories approach for speaker recognition is extremely useful as it allows us to study which phonetic units of speech are robust to voice disguise. This in turn can be useful for applying some sort of weighting function to the phonemes prior to recognition. Another key advantage of using phonetic categories approaches is that they allow for recognition even from small duration of speech frames (typically 25 ms).

It is also important to mention here that we assume that only the natural voice samples are available to train the speaker recognition systems. This is due to the fact that in any real application, it would be almost impossible to know the disguised voices of a speaker beforehand. Different speakers also possess different abilities to disguise their voice, thus a speaker recognition should be generalizable making it robust to attacks of different kinds of disguise from various speakers. This fact also makes it extremely challenging for traditional approaches, which disregard phonetic information, to accurately detect speakers from their disguised voices. The phonetic categories approaches, on the other, can rely on disguise-robust phonemes, i.e. phonemes which have some what similar characteristics for the disguised and natural voice of a speaker. Thus, the phonetic categories approach makes it possible to model fine-grained phonetic information which can be useful to identify speakers more reliably from both their disguised and natural voices.
Glottal and Vocal Tract Analysis of Voice Impersonations

Voice impersonators possess a flexible voice and thus can change their voice identity. They are able to imitate various people and characters which differ in age, gender, accent and voice quality. To understand why human impersonators are successful and what parameters they rely on to change their voice, we first analyze nine voices produced by a professional voice impersonator. The different acoustical measures are computed and their linguistic implications are discussed in this chapter. The acoustical measures include pitch, speech rate and formant frequencies. We then generalize the findings by extending the analysis to include three different impersonators. To understand how the impersonators make use of their vocal folds, we also analyze the vocal fold properties through EGG signals. Finally, we evaluate the success of impersonators in deceiving humans through a subjective listening test and propose an objective measure which correlates well with the subjective ratings of voice disguise.

The rest of this chapter is organized as follows: In Section 3.1, we present the acoustic and linguistic analysis of a single impersonator, Section 3.2 describes the analysis for three impersonators and finally Section 3.3 concludes the chapter.

3.1 Analysis of a single impersonator

Here we perform an acoustic and linguistic analysis of the various impersonations of a single impersonator. The parameters we have measured and analyzed include fundamental frequency (pitch), speaking rate (speech rate) and vowel formant frequencies (F1, F2 and F3). Pitch and speaking rate are prosodic features
Table 3.1: Abbreviations of the nine different voices from the voice artist.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPF</td>
<td>High Pitched Female</td>
</tr>
<tr>
<td>NF</td>
<td>Nasal Female</td>
</tr>
<tr>
<td>OF</td>
<td>Old Female</td>
</tr>
<tr>
<td>OM</td>
<td>Old Male</td>
</tr>
<tr>
<td>N</td>
<td>Natural</td>
</tr>
<tr>
<td>SOF</td>
<td>Scratchy Old Female</td>
</tr>
<tr>
<td>YF</td>
<td>Young Female</td>
</tr>
<tr>
<td>YG</td>
<td>Young Girl</td>
</tr>
<tr>
<td>YM</td>
<td>Young Male</td>
</tr>
</tbody>
</table>

of the speech signal that may be both speaker-specific and relevant for conveying meaning. Formant frequencies are the primary acoustic correlates of differences between vowels and are primarily determined by shape characteristics of the vocal tract. We also show how these parameters might be affected by language or dialect specific features such as those associated with regional accent.

3.1.1 Speech Data

The speech data was collected from a professional voice artist, who served as the impersonator in our study. She is a middle-aged female whose first language is English. The speech data consists of recordings of a single sentence by the voice artist in nine different voices as listed in Table 3.1. The following sentence which is from the VOICES 1.0 database [Kain, 2001] was used: *To further his prestige, he occasionally reads the Wall Street Journal.* Each voice represents an impersonation of a distinct (fictional) character by the voice artist, differing from each other in terms of age, gender and regional identity. The voice artist used a relatively neutral variety of English (resembling many British influenced Asian standard varieties) for all the voices except NF and SOF, which were perceived by a trained phonetician as having a markedly North American quality. Two distinct voice qualities, nasalization and creak, were also utilized by the voice artist for NF and SOF respectively. All the voices were intended to be emotionally neutral, and this impression was confirmed by subsequent inspection.

The recordings were done in a professional home studio using a Shure PG98 Dynamic Vocal microphone with Edirol 24 bit / 96 kHz USB Audio capture and Sonar LE software package for audio processing. The speech data was originally recorded in MP3 format at a sampling rate of 44 kHz using 16 bits/sample. It was later converted to mono WAV format and downsampled to 16 kHz using Audacity®.
3.1.2 Results

The speech data was analyzed for three acoustic parameters: fundamental frequency, speech rate and formant frequencies. Praat [Boersma and Weenink, Retrieved October 21, 2011] was used to extract the fundamental frequency and the formant frequencies. The fundamental frequency was calculated every 10 milliseconds which is the default step size in Praat while its range was set to 75-600 Hz. The speech rate was calculated using the algorithm in [De Jong and Wempe, 2009] which computes the number of intensity maxima over the entire duration of the speech signal to give an estimate of the number of syllables. The speech rate is thus expressed in terms of the number of syllables per second. The formant frequencies were calculated every 6.25 milliseconds with the maximum formant frequency set to 5500 Hz during the analysis. The number of poles were set to 12. The speech data was segmented at both the word and phone level using the forced alignment algorithm from the HTK Speech Recognition Toolkit [Young et al., 2006]. The CMU dictionary [Weide, 1998] was used to provide the phones for each word. The results of the automated segmentation were then manually corrected by a trained phonetician.

3.1.2.1 Fundamental frequency (F0)

Fundamental frequency is the acoustic correlate of perceived pitch in speech. Since F0 characteristics of speech may vary significantly from speaker to speaker, it is important to consider their relevance for voice identity. The mean fundamental frequency depends largely on the size of the vocal folds. In general, men have lower values of mean F0 compared to women since they have larger vocal folds [Latinus and Belin, 2011]. It is therefore interesting to see how the impersonator changes her F0 mean and standard deviation with respect to the nine voices. Figure 3.1 shows the mean and the standard deviation of F0 for all the voices. From the graph, it can be seen that the voice artist is flexible with her mean F0 and can vary it within a range of 132 to 260 Hz. The voice artist’s natural voice N has a mean F0 of 165 Hz which is rather low for an English speaking female [Brown et al., 1991].

From Figure 3.1, it is also observed that the standard deviation of F0 for the male voices, i.e., OM and YM, is smaller compared to the female voices. As a result, the male voices sound more monotonous. Since women tend to vary their F0 more than men [van Rie and van Bezooijen, 1995, Brend, 1975], varying the F0 may be one strategy for feminizing the voice. These differences in the variation of pitch for male and female voices can also be observed from the pitch contour plots. The pitch contour plots represent the evolution of the perceived

---

1A phone in phonetics is a speech segment which possesses distinct physical and perceptual properties.
3.1. ANALYSIS OF A SINGLE IMPERSONATOR

pitch of the sound over time. Figure 3.2 shows the pitch contour for HPF and YM. It can be observed that HPF has more peaks and valleys compared to YM, particularly after 1.5 seconds. Additionally, the difference in the peaks and valleys is more pronounced for HPF reflecting a larger F0 range. These differences in the pitch contour clearly show that the voice artist varies her intonation patterns when impersonating male and female voices. Impressionistically, it can be noted that the overall effect is that HPF sounds more expressive than YM.

It is also worth mentioning here that intonation, or variations in pitch height and amplitude that are temporally aligned to the segmental description, is an important parameter for the production of natural sounding speech. However, for many speech processing applications, such as, voice conversions [Machado and Queiroz, 2010, Erro et al., 2007] the pitch contour of the source speaker $F_0_s$ is transformed to the pitch contour of the target speaker $F_0_t$ according to Equation (3.1) which is defined as
3.1. ANALYSIS OF A SINGLE IMPERSONATOR

\[
F_0_t = \mu_t + \frac{\sigma_t}{\sigma_s} (F_0_s - \mu_s) \tag{3.1}
\]

where \( \mu_s, \sigma_s, \mu_t, \sigma_t \) represent the mean and standard deviation of the F0 for the source and target speakers respectively.

The linear transformation in Equation (3.1) fails to capture finer intonational details of the target speaker and also cannot model the local changes in F0. As a result, the converted voice often lacks naturalness specially when converting from male to female voice [Machado and Queiroz, 2010]. Some methods [Helander and Nurminen, 2007, Inanoglu, 2003] have started to include detailed prosody modeling in voice conversion systems but there is still ample room for improvement. The results indicate that intonation is also an important parameter used by the voice artist for projecting different voice identities. More detailed models of intonation are thus required which can be useful for speaker modeling.

Figure 3.2: Pitch contours for HPF (solid line) and YM (bold line) calculated using Praat.
3.1. ANALYSIS OF A SINGLE IMPERSONATOR

3.1.2.2 Speech Rate

Our findings show that the speech rate is also an important cue for voice identity. It reflects among other factors the speaking style of an individual. The effect of gender and age on speaking rate has been investigated before. It has been shown that men generally speak faster than women [Byrd, 1994, Yuan et al., 2006, Jacewicz et al., 2009] while young adults tend to speak faster than older adults [Yuan et al., 2006, Jacewicz et al., 2009, Smith et al., 1987]. Speech rate was therefore examined in order to understand how the speaker exploits changes in her articulation rate to achieve different voice identities. Figure 3.3 shows the speech rate for the nine voices. It is observed that the male voices YM and OM have a higher speaking rate compared to the female voices. It can also be observed from Figure 3.3 that the speech rate of the voice artist’s natural voice N is closer to all the female voices except NF, YF and YG. The YG voice has the highest speaking rate among all the female voices.

These results suggest that the speech rate is an important parameter used by the voice artist for impersonating different ages and genders. However, for applications such as voice conversion a typical method is to employ a time scale modification approach on a frame by frame basis. This is done to adapt the local speaking rate of the source speaker to that of the target speaker. Since, these methods operate on a frame level, the speaking style is often not natural [Leutelt and Heute, 2002]. Therefore, further research is needed to build on these methods in a way that takes linguistic information into account for modeling the speech rate of speakers.

3.1.2.3 Formant frequencies

Formant frequencies are identified by the peaks in the spectral envelope of the speech signal, and are determined by the natural resonances of the vocal tract. For a given speaker, changes in formant frequencies depend primarily on changes in the shape and position of the articulators (tongue, lips, jaw, etc.) during speech production. For linguistic purposes, the first three formant frequencies, F1, F2 and F3, are the principal acoustic correlates of perceptual differences among vowel categories, and are also responsible for subtle differences between vowel tokens within a category. Crucially, formant values also depend inversely on vocal tract length. In general, men have an average vocal tract length of about 20 cm while female vocal tracts are 15-20% shorter in length [Fant, 1966]. Therefore it is expected that men have lower overall formant frequencies than females [Peterson and Barney, 1952] when producing the same vowel. Given that formant frequencies can be an important cue to differences between speakers, they are predicted to be an important feature for voice identity [Coleman, 1976]. In order to understand how the voice artist exploits changes in her vocal tract while impersonating, we have analyzed two vowels here, specifically, /ɔː/ and /ʊə/. It is found that F1 and
F2 of the stressed vowel /ɔː/ in *journal* and the stressed monophthong /ʌ:/ in *wall* clearly demonstrate the strategies employed by the voice artist to change her identity. This is illustrated in Figure 3.4 which shows the F1 vs F2 plot for the two vowels /ɔː/ and /ʌ:/ in *wall*. It is observed that the male voices: OM and YM have a lower value of F1 compared to the female voices. The natural voice of the voice artist N is closer to the SOF and OF voices while the young female YF has a higher F1. However, the NF and HPF voices do seem to change position with the vowel. The YG has a particularly high value of F2. These results suggest that the voice artist is able to alter her perceived vocal tract length while impersonating different ages, genders and voice qualities.

Linguistically, these two vowels are somewhat special since they are essentially key markers of regional accent. For example, /ɔː/ is pronounced with /r/-coloring, or rhotacization, in only some of the voices, namely NF and SOF. The primary acoustic correlate of such rhotacization is a lowering of F3, though it may also be accompanied by slight raising of F2 [Stevens, 2000]. This effect can be
3.1. ANALYSIS OF A SINGLE IMPERSONATOR

Figure 3.4: F1 vs F2 for two different vowels
3.1. ANALYSIS OF A SINGLE IMPERSONATOR

3.1.1. FORMANT FREQUENCIES AND VOWEL IDENTIFICATION

Figure 3.5: The first three formants: F1, F2, F3 for /ɜ:/ in journal for the nine voices.

observed in Figure 3.5 which shows the first three formant frequencies for /ɜ:/ in journal. Perceptually, the overall impression of F3-lowering is that the speaker is using a North American dialect of English. At least part of the discriminability seen in Figure 3.4(a), therefore, may be a result of the speaker’s ability to access different regional accents, rather than from voice quality per se. The other vowel /ɔ:/, shown in Figure 3.4(b), is similar, in that the degree of lip rounding, manifested acoustically as a lowering of F2, is highly associated with differences in regional accent. Still, in the case of /ɔ:/, it is possible that the voice artist may have been exploiting changes in F2 via lip rounding to achieve speaker-specific effects that are unrelated to regional accent such as vocal tract length or cross-sectional area.

Together, these results show that the high degree of variability exhibited in the voice artist’s production of vowels is clearly a major resource that she exploits to achieve different voice identities.
3.1.3 Conclusion

In this Section, an acoustic and linguistic analysis of nine different voices produced by a voice artist was performed. From auditory analysis, it is clear that the voice impersonator is successful in changing her voice identity in terms of age, gender and voice quality. The results also confirm that these changes are accompanied by changes in mean F0, F0 dispersion, intonation pattern, speaking rate and vocal tract shape to achieve different voice identities. It is concluded that the acoustical measures are greatly affected by language-dependent features such as those associated with regional accent.

Together, these findings suggest that it is possible to synthetically generate a range of distinct voice identities by systematically recombining a finite number of linguistic and non-linguistic features of the speech signal. Further research is needed to determine whether the set of factors considered in our study are sufficient for generating voices that rate highly in terms of individuality and naturalness, or whether more features must be incorporated.

3.2 Analysis of three impersonators

In the previous section, we analyzed the nine different voices from a single voice impersonator using a single sentence. In this section, we build upon the previous analysis and better generalize our earlier findings by (i) using three impersonators (including the one female impersonator from Section 3.1), (ii) using a more comprehensive sample of vowel categories for the analysis of the vocal tract characteristics, and (iii) using a total of 486 sentence tokens for analysis (versus 9 in the previous section).

3.2.1 Data Collection

Three professional voice artists (one female, two male) served as the impersonators in this study. We refer to them henceforth as impersonator 1F, 2M and 3M respectively. The first and dominant language of all three impersonators is English, with some differences in dialectal features (South Asian, Southeast Asian and North American for 1F, 2M and 3M, respectively).

Data collection took place inside a sound-attenuated room, and synchronous speech and EGG signals were recorded from the productions. The speech signal was recorded using an AKG C520L head-mounted condenser microphone. The EGG signals were obtained using a EG2-PCX2 Electroglottogram by Glottal Enterprises [Rothenberg, 1992]. This required placing two electrodes, 35mm

\[2\] The database of voice impersonators is in the process of licensing and will be available soon for the research community.
3.2. ANALYSIS OF THREE IMPERSONATORS

in diameter, externally over the larynx in order to measure the electrical conductance of the vocal folds during voiced phonation. The analog speech and EGG signals were captured on separate channels using a Zoom H4n recorder, and were digitized in WAV format at a sampling rate of 44.1 kHz with 16-bit resolution. Following the recording, the speech data was segmented at both the word- and phone-level using the Penn Phonetics Lab Forced Aligner [Yuan and Liberman, 2008]. The results of the automated segmentation were then manually corrected by a trained phonetician.

Protocol

The impersonators were given no target speakers to imitate and had the freedom to choose the voices they wanted to impersonate. They were instructed to use a consistent regional variety of English across the nine voices being impersonated, but were given freedom to vary any other identity characteristics of the voices including age and gender. Each voice impersonator produced nine distinct voice identities which included eight (fictional) character voices and their natural voices. Thus a total of 27 distinct voice identities were produced by the voice impersonators. All of these voices were natural-sounding and readily distinguished from each other.

Speech material

The same speech materials were used for all 27 voice samples, and consisted of nine short sentences. For each impersonated voice, the impersonators produced the nine sentences in a sequence, and then repeated the sequence in the same voice, for a total of 18 sample sentences per voice. Thus a total of 486 sentences were collected for analysis. For each vowel in each voice, there were about 5-6 tokens as shown in Figure 3.13. The durations of the different phones are shown in Table 5.1.

Each sentence included two monosyllabic target words containing one of the vowels /æ/, /ɛ/, /i/, /u/, and /E/. These target vowels were chosen (a) to provide a representative sample of the overall vowel ‘space’ of English (i.e., the organization of vowels in the F1-F2 plane, explained in more detail in Section 3.2.2.3), and (b) because they are relatively robust to subtle differences in regional dialect (e.g., the vowel in ‘heard’ was excluded on this basis, since American speakers tend to produce it with a stronger ‘r’-quality than most British speakers). Factors affecting word prominence, such as sentence stress and phrasing, are known to affect vowel formant measures [Lee and Cole, 2006]. To maximize consistency across samples, therefore, target words were placed in positions within the sentence that are associated with maximal prominence. Specifically, the sentences were designed so that target words would be produced with a nuclear accent and
occur at the right edge of an intonational phrase boundary. The list of sentences can be found in Appendix A.

### 3.2.2 Analysis and Results

From the analysis of the voice artist presented in Section 3.1, we have established the importance of the glottal (F0, speech rate) and vocal tract (F1, F2, F3) measures. In this section, we extend the analysis to include three artists. We also provide an analysis of the vocal fold properties through the use of electroglottograph signals.

#### 3.2.2.1 Glottal measures

**a) F0**

Fundamental frequency (F0) is the acoustic correlate of perceived pitch in speech. Since certain F0 characteristics of speech may vary significantly from speaker to speaker, it is important to consider their relevance for voice identity. A number of studies [Hillenbrand and Clark, 2009, Honorof and Whalen, 2010, Sambur, 1975] have investigated the role of mean F0 values for distinguishing the voices of men and women. Overall, the mean F0 tends to be inversely correlated with the length and size of the vocal folds, thus men generally have a lower mean F0 compared to women [Latinus and Belin, 2011], while adults tend to have a lower mean F0 than children. Additionally, women tend to exhibit a higher degree of temporal variation in F0 than men [van Rie and van Bezooijen, 1995, Brend, 1975], meaning that there are more frequent peaks and valleys in the temporal F0 contour, and the differences between those peaks and valleys tend to be larger. It is therefore important to consider the extent to which the voice impersonators exploit this variability in F0 characteristics in creating the various voice identities.

For the F0 analyses, Praat [Boersma and Weenink, Retrieved October 21, 2011] was used to first obtain F0 samples at 10 ms intervals, using a frequency window of 75-600 Hz. The mean and standard deviation were estimated from all samples occurring within the voiced portions of all 18 sentences for a given voice. Table-3.2 shows the mean and the standard deviation of F0 for all the voices of the three impersonators arranged by mean F0. The various voice identities are represented by $V_i$, where $i$ refers to the voice number for that speaker, and $V_1$ is always the natural voice and is indicated in bold. In some cases, the impersonators provided labels for the voices that were indicative of either age or gender identity. We indicate this using a combination of the labels “Y” (young), “O” (old), “M” (male), and “F” (female). Speakers chose their voices freely, and numbering was assigned arbitrarily, so there is no correspondence between same-numbered voices across impersonators.
3.2. ANALYSIS OF THREE IMPERSONATORS

Table 3.2: Mean and standard deviation of F0 in Hertz. The natural voices are shown in bold.

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>μ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>V₈</td>
<td>YM</td>
<td>151.43</td>
<td>33.78</td>
</tr>
<tr>
<td>V₅</td>
<td>OM</td>
<td>159.39</td>
<td>47.28</td>
</tr>
<tr>
<td>V₁</td>
<td>-</td>
<td>196.75</td>
<td>53.81</td>
</tr>
<tr>
<td>V₆</td>
<td>OF</td>
<td>212.72</td>
<td>64.15</td>
</tr>
<tr>
<td>V₃</td>
<td>-</td>
<td>258.80</td>
<td>88.27</td>
</tr>
<tr>
<td>V₄</td>
<td>YM</td>
<td>266.96</td>
<td>63.66</td>
</tr>
<tr>
<td>V₉</td>
<td>YF</td>
<td>274.73</td>
<td>67.81</td>
</tr>
<tr>
<td>V₇</td>
<td>YF</td>
<td>294.34</td>
<td>70.99</td>
</tr>
<tr>
<td>V₂</td>
<td>-</td>
<td>414.87</td>
<td>88.15</td>
</tr>
</tbody>
</table>

(a) Impersonator 1F

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>μ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>V₁</td>
<td>-</td>
<td>94.52</td>
<td>25.06</td>
</tr>
<tr>
<td>V₆</td>
<td>-</td>
<td>100.49</td>
<td>36.59</td>
</tr>
<tr>
<td>V₂</td>
<td>-</td>
<td>103.82</td>
<td>44.53</td>
</tr>
<tr>
<td>V₅</td>
<td>-</td>
<td>124.45</td>
<td>28.66</td>
</tr>
<tr>
<td>V₇</td>
<td>OM</td>
<td>150.53</td>
<td>34.63</td>
</tr>
<tr>
<td>V₃</td>
<td>-</td>
<td>155.73</td>
<td>51.39</td>
</tr>
<tr>
<td>V₉</td>
<td>YM</td>
<td>170.39</td>
<td>57.79</td>
</tr>
<tr>
<td>V₈</td>
<td>OM</td>
<td>204.14</td>
<td>58.27</td>
</tr>
<tr>
<td>V₄</td>
<td>YF</td>
<td>235.85</td>
<td>62.81</td>
</tr>
</tbody>
</table>

(b) Impersonator 2M

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>μ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>V₈</td>
<td>-</td>
<td>113.73</td>
<td>52.98</td>
</tr>
<tr>
<td>V₇</td>
<td>OM</td>
<td>123.75</td>
<td>36.36</td>
</tr>
<tr>
<td>V₁</td>
<td>-</td>
<td>125.25</td>
<td>32.60</td>
</tr>
<tr>
<td>V₆</td>
<td>-</td>
<td>142.54</td>
<td>36.92</td>
</tr>
<tr>
<td>V₉</td>
<td>-</td>
<td>149.67</td>
<td>48.30</td>
</tr>
<tr>
<td>V₄</td>
<td>-</td>
<td>166.93</td>
<td>39.94</td>
</tr>
<tr>
<td>V₂</td>
<td>-</td>
<td>178.53</td>
<td>35.82</td>
</tr>
<tr>
<td>V₅</td>
<td>-</td>
<td>185.98</td>
<td>40.43</td>
</tr>
<tr>
<td>V₃</td>
<td>-</td>
<td>309.19</td>
<td>70.24</td>
</tr>
</tbody>
</table>

(c) Impersonator 3M
A first glance at Table 3.2 reveals that all impersonators were flexible with their pitch in creating different voice identities. The mean F0 exhibited a range of at least one octave across the voices for each impersonator. Not surprisingly, the male voices ranked consistently lower than the female voices both in terms of mean F0 and standard deviation. For the female speaker 1F, the two voices with the lowest mean and standard deviation (V5 and V8) are both male, while for the male speaker 2M, the voice with the highest mean and standard deviation i.e. V4 is the only female voice he produced. While the role of age is less apparent, it can be noted that for 1F, the “old female” (V6) has the lowest mean and standard deviation among the female voices, while the “old male” is very close to the bottom of the range. 2M and 3M show a similar tendency. Figure 3.6 plots the old (O) and young (Y) voices of all three impersonators together. From this figure, it can be observed that in general young voices have a higher mean F0 compared to the old voices. Similarly, a plot of the male (M) and female (F) voices in Figure 3.7 reveals that female voices generally have higher mean F0 than male voices.

A one-way ANOVA confirmed that the effect of voice on mean F0 is signifi-
3.2. ANALYSIS OF THREE IMPERSONATORS

Figure 3.7: The F0 for the male and female voices of all three impersonators.

The F0 for all three impersonators ($F^3(8, 51216)=7606.244, F(8, 42224)=4755.150, F(8, 43253)=8367.154; p<0.05$ for 1F, 2M and 3M, respectively).

These results confirm our assumption that the impersonators would exploit the stereotypical correspondences between F0 and identity in order to achieve different voice identities. It also illustrates the sense in which variability for a given parameter may be limited on a speaker-specific basis. Even when 1F was using a stereotypically male voice, her mean F0 was higher than the lowest voices for the two male impersonators, 2M and 3M. Similarly, neither 2M or 3M exhibited a mean F0 as high as the maximum for 1F, and their standard deviations were remarkably consistent in being lower than those for 1F. Interestingly, the lowest voice for 2M is his natural voice (on both measures), suggesting that he typically speaks near the bottom of his range, and can only increase both the mean and standard deviation of F0 in order to achieve variation in voice identity.

---

3 The $F$ value is a statistic for reporting ANOVA results. If the null hypothesis is true, $F$ will have a value close to 1.0 most of the time. A large $F$ value implies that the variation among group means is more than chance.
3.2. ANALYSIS OF THREE IMPERSONATORS

Table 3.3: Mean and standard deviation of the speech rate (syllables per second). The natural voices are shown in bold.

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_5$</td>
<td>OM</td>
<td>2.33</td>
<td>0.36</td>
</tr>
<tr>
<td>$V_7$</td>
<td>YF</td>
<td>2.52</td>
<td>0.35</td>
</tr>
<tr>
<td>$V_6$</td>
<td>OF</td>
<td>2.53</td>
<td>0.29</td>
</tr>
<tr>
<td>$V_3$</td>
<td>-</td>
<td>3.48</td>
<td>0.47</td>
</tr>
<tr>
<td>$V_1$</td>
<td>-</td>
<td>3.95</td>
<td>0.61</td>
</tr>
<tr>
<td>$V_2$</td>
<td>-</td>
<td>3.81</td>
<td>0.47</td>
</tr>
<tr>
<td>$V_9$</td>
<td>YF</td>
<td>4.08</td>
<td>0.60</td>
</tr>
<tr>
<td>$V_4$</td>
<td>YM</td>
<td>4.31</td>
<td>0.41</td>
</tr>
<tr>
<td>$V_8$</td>
<td>YM</td>
<td>4.48</td>
<td>0.46</td>
</tr>
</tbody>
</table>

(a) Impersonator 1F

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_8$</td>
<td>OM</td>
<td>3.19</td>
<td>0.48</td>
</tr>
<tr>
<td>$V_5$</td>
<td>-</td>
<td>3.25</td>
<td>0.45</td>
</tr>
<tr>
<td>$V_6$</td>
<td>-</td>
<td>3.31</td>
<td>0.51</td>
</tr>
<tr>
<td>$V_7$</td>
<td>OM</td>
<td>3.46</td>
<td>0.47</td>
</tr>
<tr>
<td>$V_2$</td>
<td>-</td>
<td>3.54</td>
<td>0.53</td>
</tr>
<tr>
<td>$V_3$</td>
<td>-</td>
<td>3.94</td>
<td>0.55</td>
</tr>
<tr>
<td>$V_1$</td>
<td>-</td>
<td>4.17</td>
<td>0.48</td>
</tr>
<tr>
<td>$V_4$</td>
<td>YF</td>
<td>4.19</td>
<td>0.57</td>
</tr>
<tr>
<td>$V_9$</td>
<td>YM</td>
<td>4.22</td>
<td>0.55</td>
</tr>
</tbody>
</table>

(b) Impersonator 2M

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_7$</td>
<td>OM</td>
<td>3.42</td>
<td>0.55</td>
</tr>
<tr>
<td>$V_1$</td>
<td>-</td>
<td>3.70</td>
<td>0.52</td>
</tr>
<tr>
<td>$V_5$</td>
<td>-</td>
<td>3.85</td>
<td>0.64</td>
</tr>
<tr>
<td>$V_9$</td>
<td>-</td>
<td>4.03</td>
<td>0.49</td>
</tr>
<tr>
<td>$V_6$</td>
<td>-</td>
<td>4.06</td>
<td>0.51</td>
</tr>
<tr>
<td>$V_3$</td>
<td>-</td>
<td>4.09</td>
<td>0.65</td>
</tr>
<tr>
<td>$V_4$</td>
<td>-</td>
<td>4.20</td>
<td>0.71</td>
</tr>
<tr>
<td>$V_2$</td>
<td>-</td>
<td>4.53</td>
<td>0.69</td>
</tr>
<tr>
<td>$V_8$</td>
<td>-</td>
<td>4.61</td>
<td>0.75</td>
</tr>
</tbody>
</table>

(c) Impersonator 3M
3.2. ANALYSIS OF THREE IMPERSONATORS

b) Speech Rate

Speech rate has been linked to a number of stylistic factors, though it can also be related to speaker identity features, including gender and age. Men, for example, generally speak faster than women [Byrd, 1994, Yuan et al., 2006, Jacewicz et al., 2009] while young adults tend to speak faster than older adults [Yuan et al., 2006, Jacewicz et al., 2009, Smith et al., 1987]. We therefore explored the extent to which differences in speech rate were exploited by the voice impersonators in creating different voice identities.

The speech rate, in syllables per second, was calculated for each voice by counting the total number of syllables in each sample and then dividing by the overall duration of all non-silent portions of the sample. Table-3.3 shows the average speech rate and standard deviation for the voices of each impersonator arranged by the average speech rate. All impersonators showed differences in speech rate across the voices of at least 32% (for 2M) and as much as 92% (for 1F). Consistent with earlier studies on age effects, the highest and lowest speaking rates
for each impersonator were instantiated by “young” and “old” voices respectively. Additionally, “young” and “old” voices tend to cluster at the top and bottom of the range, respectively, for each impersonator. The exception is V₇ of 1F, which impressionistically sounds like a small child speaking deliberately and somewhat effortfully. This is more clearly illustrated in Figure 3.8 where the speech rate for the young (Y) and old (O) voices has been plotted collectively for all three impersonators.

The role of gender, however, is less clear as evident from Figure 3.9 which plots the speech rate of all impersonators in terms of their male (M) and female (F) voices. The fastest speaking rate for both 1F and 2M was instantiated by a male voice rather than a female one (as predicted), though overall, the effect of age appears to dominate. Since these labels do not represent controlled variables in the proper sense (e.g., a given “young” voice may not correspond to precisely the same age as another “young” voice), it is not possible to clearly isolate the contribution of gender. Nevertheless, our results confirm the prediction that im-
personators use speech rate as an important parameter in the creation of distinct voice identities, and they provide an indication of the amount of variability that is achievable for a given speaker within the bounds of naturalness and individual physical traits. A one-way ANOVA revealed that there was a significant effect of voice on the speech rate for all three impersonators ($F(8, 153)=53.760$, $F(8, 153)=12.800$, $F(8, 153)=6.603$; $p<0.05$ for 1F, 2M and 3M, respectively).

### 3.2.2.2 Measures using the Electroglottograph

The Electroglottograph (EGG) signal provides an estimate of the vocal fold contact area [Childers and Krishnamurthy, 1985] by measuring the electrical conductance between the vocal folds. It is useful for analyzing the complex three dimensional movements of the vocal folds since it provides an image of the signal generated at the glottis. Compared to the speech signal, then, the EGG signal is generally free from the filtering effects of the vocal tract. Historically, the EGG signal has been used for detecting voice quality [Fourcin, 2000] as well as for speaker identification [Campbell et al., 2003]. Some studies including [Neocleous and Naylor, 1998] have suggested that speakers do not possess as much voluntary control over their vocal fold behavior as compared to their vocal tract characteristics. The rationale behind using the EGG signals is therefore to explore whether and in what ways the voice impersonators actively exploit differences in vocal fold patterns while impersonating different voices. The measure used for our study was the Open Quotient (OQ), which is directly related to the timing characteristics of the vocal folds, and is described in detail below. A number of studies have reported a correspondence between this measure and various identity features, including age and gender [Higgins and Saxman, 1991, Winkler and Sendlmeier, 2005, Ma and Love, 2010], as well as voice quality [Fourcin, 2000]. On that basis, we predicted that OQ would differ across voices for a given speaker, and that these differences would show an approximate correspondence with the identity labels provided by the impersonators. For voiced phonation, the vocal folds vibrate in a periodic manner, moving in and out of contact with each other. Thus, the EGG signal also varies periodically as a function of the contact area between the vocal folds. Now consider a vocal fold vibratory cycle in which the vocal folds are initially not in contact, resulting in the electrical conductance being minimum. As the vocal folds begin to move in contact, the electrical conductance starts to increase. The time instant at which the glottis becomes closed is called the Glottal Closing Instant (GCI). The glottal closing is generally abrupt and appears as a steep slope in the EGG signal as shown in Figure 3.10. It is widely accepted that the GCI appears as a sharp positive peak in the Differentiated ElectroGlottograph (DEGG) signal [Childers and Krishnamurthy, 1985, Henrich et al., 2004]. The glottis then remains closed for a short period of time before the vocal folds start separating again, causing the measured electrical conductance to decrease.
3.2. ANALYSIS OF THREE IMPERSONATORS

The time instant at which the glottis becomes opened is called the Glottal Opening Instant (GOI). The GOI appears as a low amplitude peak in the DEGG signal with a polarity opposite to that of the GCI peak [Henrich et al., 2004]. The EGG and DEGG signals corresponding to a voiced segment of speech together with the labeled GCIs and GOIs are shown in Figure 3.10.

Using the GCIs and GOIs as two distinct landmarks in the DEGG signal, we can now define some of the glottal parameters, i.e., the open and close phase as shown in Figure 3.10. The period of time for which the glottis remains closed over a glottal cycle is called the Closed Phase (CP). For the $k^{th}$ glottal period $T(k)$, the CP$(k)$ and OP$(k)$ are defined as

$$\text{CP}(k) = \text{GOI}(k) - \text{GCI}(k).$$

(3.2)

The period of time for which the glottis remains opened over a glottal cycle is called the Open Phase (OP). For the $k^{th}$ glottal period $T(k)$, the OP$(k)$ is given as

$$\text{OP}(k) = \text{GCI}(k + 1) - \text{GOI}(k).$$

(3.3)
3.2. ANALYSIS OF THREE IMPERSONATORS

The time period of the $k^{th}$ glottal cycle is then defined as

$$T(k) = CP(k) + OP(k).$$

(3.4)

Once the CP and OP are obtained, we can define the open quotient.

**a) Open Quotient (OQ)**

The OQ represents the percentage of time for which the glottis remains opened over a glottal period. For the $k^{th}$ glottal period, the OQ($k$) is defined as

$$OQ(k) = OP(k)/T(k).$$

(3.5)

Various studies have investigated the relationship between the OQ and the perceived age and gender of the speaker. The authors in [Higgins and Saxman, 1991, Winkler and Sendlmeier, 2005] report that the OQ decreases with increasing age.
### 3.2. ANALYSIS OF THREE IMPERSONATORS

Table 3.4: Mean and standard deviation of open quotient. The natural voices are shown in bold.

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_5$</td>
<td>OM</td>
<td>0.60</td>
<td>0.11</td>
</tr>
<tr>
<td>$V_6$</td>
<td>OF</td>
<td>0.60</td>
<td>0.20</td>
</tr>
<tr>
<td>$V_8$</td>
<td>YM</td>
<td>0.61</td>
<td>0.13</td>
</tr>
<tr>
<td>$V_3$</td>
<td>-</td>
<td>0.63</td>
<td>0.22</td>
</tr>
<tr>
<td>$V_4$</td>
<td>YM</td>
<td>0.70</td>
<td>0.13</td>
</tr>
<tr>
<td>$V_7$</td>
<td>YF</td>
<td>0.72</td>
<td>0.12</td>
</tr>
<tr>
<td>$V_1$</td>
<td>-</td>
<td>0.72</td>
<td>0.12</td>
</tr>
<tr>
<td>$V_9$</td>
<td>YF</td>
<td>0.73</td>
<td>0.09</td>
</tr>
<tr>
<td>$V_2$</td>
<td>-</td>
<td>0.78</td>
<td>0.09</td>
</tr>
</tbody>
</table>

(a) Impersonator 1F

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_1$</td>
<td>-</td>
<td>0.54</td>
<td>0.14</td>
</tr>
<tr>
<td>$V_6$</td>
<td>-</td>
<td>0.57</td>
<td>0.16</td>
</tr>
<tr>
<td>$V_8$</td>
<td>OM</td>
<td>0.59</td>
<td>0.17</td>
</tr>
<tr>
<td>$V_7$</td>
<td>OM</td>
<td>0.61</td>
<td>0.13</td>
</tr>
<tr>
<td>$V_5$</td>
<td>-</td>
<td>0.65</td>
<td>0.15</td>
</tr>
<tr>
<td>$V_9$</td>
<td>YM</td>
<td>0.75</td>
<td>0.08</td>
</tr>
<tr>
<td>$V_2$</td>
<td>-</td>
<td>0.78</td>
<td>0.12</td>
</tr>
<tr>
<td>$V_3$</td>
<td>-</td>
<td>0.79</td>
<td>0.07</td>
</tr>
<tr>
<td>$V_4$</td>
<td>YF</td>
<td>0.79</td>
<td>0.07</td>
</tr>
</tbody>
</table>

(b) Impersonator 2M

<table>
<thead>
<tr>
<th>Voice</th>
<th>Label</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_9$</td>
<td>-</td>
<td>0.43</td>
<td>0.09</td>
</tr>
<tr>
<td>$V_4$</td>
<td>-</td>
<td>0.46</td>
<td>0.07</td>
</tr>
<tr>
<td>$V_5$</td>
<td>-</td>
<td>0.48</td>
<td>0.05</td>
</tr>
<tr>
<td>$V_2$</td>
<td>-</td>
<td>0.48</td>
<td>0.06</td>
</tr>
<tr>
<td>$V_1$</td>
<td>-</td>
<td>0.49</td>
<td>0.09</td>
</tr>
<tr>
<td>$V_7$</td>
<td>OM</td>
<td>0.52</td>
<td>0.17</td>
</tr>
<tr>
<td>$V_6$</td>
<td>-</td>
<td>0.55</td>
<td>0.14</td>
</tr>
<tr>
<td>$V_8$</td>
<td>-</td>
<td>0.58</td>
<td>0.23</td>
</tr>
<tr>
<td>$V_3$</td>
<td>-</td>
<td>0.83</td>
<td>0.07</td>
</tr>
</tbody>
</table>

(c) Impersonator 3M
3.2. ANALYSIS OF THREE IMPERSONATORS

Figure 3.12: The open quotient for the the male and female voices of all three impersonators.

For females, while in [Higgins and Saxman, 1991] the OQ increased with increasing age for males. Since the OQ measure has been found to be mostly independent of the vowel category [Epstein, 2002], in our study, all target vowels for a given voice were used to obtain the final estimate of its mean and standard deviation. Table-3.4 shows the mean OQ and its standard deviation for the voices of all three impersonators arranged by mean OQ. The GCIs and GOIs were estimated from the DEGG signal using a newly proposed novel method which will be presented in Chapter 6.

From Table-3.4, we find that overall, the “young” voices consistently showed higher OQ means than the “old” voices (in the case of 1F and 2M). This can also be observed in Figure 3.11 which plots the OQ for the young and old voices of all impersonators. Consistent with earlier studies showing an interaction between age and gender, within the young voices, female voices had higher OQ means than male voices, while there was no difference between genders for the two old voices produced by 1F. Figure 3.12 plots the OQ for the male and female voices of all
3.2. ANALYSIS OF THREE IMPERSONATORS

Impersonators. Based on the gender alone, no clear patterns can be observed from Figure 3.12.

There was a significant effect of voice on the mean OQ for all three impersonators ($F(8, 7286)=203.998, F(8, 5257)=319.113, F(8, 6332)=1806.637; p<0.05$ for 1F, 2M and 3M, respectively). Together, these results show that the impersonators not only have significant voluntary control over their vocal fold patterns, but that they actively manipulated these patterns in order to achieve distinct voice identities.

3.2.2.3 Vocal tract measures

Formant frequencies are identified by the peaks in the spectral envelope of the speech signal, and are determined by the natural resonances of the vocal tract. For a given speaker, changes in formant frequencies depend primarily on changes in the shape and position of the articulators (tongue, lips, jaw, etc.) during speech production. For linguistic purposes, the first two formant frequencies, $F_1$ and $F_2$, are the principal acoustic correlates of perceptual differences among vowel categories, and are also responsible for subtle differences between tokens (spoken instances) of the same vowel. A useful way to visualize relationships between the vowels and formant frequencies is through a two-dimensional Cartesian plot of $F_1$ versus $F_2$, otherwise known as the “vowel space” (see Figure 3.13 for example). Each data point in the vowel space represents a token of a particular vowel category, and different vowel categories will tend to have distributions of tokens that occupy different regions of the space. The vowel category /i/, for example, tends to occupy a region corresponding to a low $F_1$ and a high $F_2$, while /a/ tends to have a high $F_1$ and a low $F_2$. It is this differentiation that makes it possible for vowel sounds, which are distributed over a continuous space, to be perceived and represented in discrete, categorical terms. Nevertheless, there is typically some overlap between the distributions of neighboring vowels, and the arrangement and positioning of vowels in the vowel space can be sensitive to dialectal [Byrd, 1994], stylistic [Moon and Lindblom, 1994], prosodic [Lee and Cole, 2006, Byrd, 1994], and importantly for our study, speaker-specific factors (esp. the length and proportioning of the vocal tract) [Stevens, 2000].

One important influence of the speaker has to do with the fact that the overall range of formant values depends inversely on vocal tract length. In general, men have a vocal tract that is approximately 15-20% longer than females [Fant, 1966], so it is expected that men have lower overall formant frequencies than females when producing the same vowel [Peterson and Barney, 1952]. Since vocal tract length increases as children grow, adults generally have lower overall formant frequencies than children. Additionally, speakers may vary in how spread out their vowels are from one another in the vowel space (otherwise known as “dispersion”). For American English speakers, this difference can be recruited
Figure 3.13: The vowel space for the nine impersonated voices of impersonator 1F. The vowel ellipses represent the 95% confidence region.
3.2. ANALYSIS OF THREE IMPERSONATORS

as a marker of identity, with female and gay male speakers generally showing higher levels of dispersion than other groups [Bradlow et al., 1996, Pierrehumbert et al., 2004]. In short, speakers tend to exhibit substantial variation in the overall positioning of vowels in the vowel space, though the relative positions of the vowels tend to be constant for a given language. Given that formant frequencies are an important cue to differences between speakers, they are predicted to be an important source of variability for voice identity construction [Coleman, 1976].

We therefore analyzed the key formants (F1 and F2) of six vowel categories in the inventory of English, in order to explore whether the voice impersonators systematically manipulated aspects of the range, positioning, and distribution of the vowels in their attempts to create distinct voice identities.

The Burg method in Praat [Boersma and Weenink, Retrieved October 21, 2011] was used to obtain formant measures by estimating the value of F1 and F2 at the temporal center of each target vowel. A frequency window of 0-5.5 kHz was used with an analysis window length of 25 ms, and the number of poles set at 12. Following extraction, a small number of tokens were identified as having potentially erroneous formant estimates based on what is typical for each vowel. Visual inspection of the Short Time Fourier Transform (STFT) time-frequency distribution (spectrogram) was then used to determine whether each such measurement was indeed erroneous, and to obtain a manual reading using Praat’s built-in formant tracking.

Figure 3.13 shows the vowel space of impersonator 1F for the six target vowels for the nine voices together with the 95% confidence regions for each of the target vowels. A confidence region is a two-dimensional generalization of a confidence interval and is represented as an ellipse placed around the point of central tendency of a distribution. For vowels, a confidence region is useful for visualizing the location and spread of a vowel category, both relative to other vowel categories, and under different conditions. The ellipses in Figure 3.13 represent the 95% confidence region for each vowel category. Points represent individual vowel tokens, and are color-coded according to vowel category. Different voices are represented by the shape of the points. This plot shows, first of all, that different vowels are subject to different kinds of variability. The vowel /i/ varies mostly in F2, for example, and /æ/ varies mostly in F1, while /e/ varies in both dimensions. Some clustering by voice is evident in Figure 3.13, suggesting that at least some of the within-vowel variation is due to the effects of voice.

To highlight this relationship, Figure 3.14 shows the same plot with some of the voices omitted for each vowel, and with the multiple tokens for each voice-vowel combination replaced by the corresponding centroid. An arrow points from the centroid for the natural voice to that for each impersonated voice. Here, it can be seen that specific voices are driving much of the variation for specific vowels. The variation in /æ/, for example, is largely driven by V5 and V7. Moreover, the same voice may affect the formants in different ways for different vowels. Voice
3.2. ANALYSIS OF THREE IMPERSONATORS

Figure 3.14: The vowel ellipses for the target vowels for some voices of impersonator 1F. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

V7, for example, is clustered in the high end of the F1 range for /i/ and /u/, but in the low end of the F1 range for /u/. Vowel tokens from the voice impersonator’s natural voice V1 generally fall close to the center of each distribution. Figure 3.15 and Figure 3.16 similarly show the natural voice for 2M and 3M, respectively, along with a selection of two other voices. Again we find that the natural voice V1 of 2M and 3M is roughly in the middle of each ellipse, and that the impersonated voices tend to deviate from the center in specific ways. The hypothesis that voices have a significant effect on the vowel formant values was confirmed by a two-factor (voice by vowel) multivariate analysis of variance (MANOVA) analysis (following [Pierrehumbert et al., 2004]) which showed that there was a significant voice and vowel interaction effect for F1 ($F(40, 270)=4.882$, partial $\eta^2=0.420$; $F(40, 270)=3.654$, partial $\eta^2=0.351$ and $F(40, 270)=5.322$, partial $\eta^2=0.441$; $p<0.05$ for 1F, 2M and 3M, respectively) and F2 ($F(40, 270)=5.902$, partial $\eta^2=0.466$; $F(40, 270)=3.925$, partial $\eta^2=0.368$ and $F(40, 270)=2.846$,}
3.2. ANALYSIS OF THREE IMPERSONATORS

Figure 3.15: The vowel ellipses for the target vowels for some voices of impersonator 2M. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

partial $\eta^2=0.297; p<0.05$ for 1F, 2M and 3M, respectively).

Crucially, the effect of each voice cannot be characterized in a general way for all vowels. It is not the case, for example, that the differences among the voices can be characterized in terms of a wholesale shift in F1 or F2 across all vowels. Nor can the differences be characterized in terms of vowel space dispersion (i.e., expansion/contraction relative to the center of the space). If that were the case, then the effect of a specific voice would be to shift all vowels universally either towards or away from the center of the space relative to the natural voice, which is clearly not the case for 3M. Instead, there is an interaction between voice and vowel category, such that the effect of a particular voice on the average formant values depends on the vowel category in question. In other words, the impersonators are making adjustments to formants on a vowel-by-vowel basis, suggesting that any successful model of either naturalness or disguise identification will at a minimum need to treat formant measures in conjunction with a linguistic parse.
3.2. ANALYSIS OF THREE IMPERSONATORS

Figure 3.16: The vowel ellipses for the target vowels for some voices of impersonator 3M. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

that includes vowel category.

A visual inspection of Figure 3.13 suggests that the natural voice tokens of each vowel are associated with lower variance (i.e., greater clustering) than those from the fictional voices. This was partly confirmed by the results of the MANOVA, which showed that the average variances in F1 across all vowels were among the lowest for the natural voices of each speaker (third, first and second lowest out of nine total voices for 1F, 2M and 3M, respectively). For F2, the average variances for the natural voices were somewhat higher (fourth, fourth and fifth out of nine total voices for 1F, 2M and 3M, respectively). Given the tilted orientation of the confidence regions for each vowel, however, we speculated that the apparent low variance of the natural voices could be captured more adequately by a measure that first decorrelates the formant measurements in the vowel space. In Section 3.2.3.2, we show that by addressing the issue of the orientation of vowel variance through Principal Component Analysis, it is possible to both capture the
intuitive sense in which natural voices are associated with lower variance in the vowel distribution, and that these distributions provide a useful metric for voice disguise identification.

Together, these results show that the high degree of variability in formants is a major resource that the impersonators exploit to achieve different voice identities. In order to capture such differences in any model, however, it is essential to take account of the implicit linguistic structure that underlies the organization of the overall vowel space. In our study, the impersonators were able to rely on their own implicit linguistic knowledge of which combinations of vowel formant frequencies are both permissible (in terms of perceptual distinctness, e.g.) and also natural-sounding. Any automated system would need to model these aspects of human linguistic competence explicitly.

### 3.2.3 Voice impersonation as a case of voice disguise

In order to understand how successful the impersonators were in their attempts to disguise their voices, we first conducted a test based on subjective human evaluations of voice disguise. This test serves as a baseline for comparison with objective measures. Then, taking inspiration from the results presented in Section 3.2.2.3, we propose a new no-reference objective metric that relies on the distributions of individual vowel categories to provide a disguise rating for a given voice. The results of the two sets of ratings are compared in order to evaluate the potential for the new objective metric in an automated voice disguise recognition system.

#### 3.2.3.1 Subjective evaluation

The subjective test obtained judgements from naive human listeners regarding whether or not a given voice sample was disguised. For this test, the entire database was divided into nine different lists, with equal numbers of samples from each impersonator, voice, and sentence. Each sentence appeared three times in a list, but never more than once by the same impersonator in the same list. The order of samples was pseudorandomized so that the minimum average distance between samples from the same impersonator was 2.0.

A total of 18 listeners participated in the study (ages 22-45, balanced bilinguals in English and one other language). Samples were presented one at a time through headphones using Psychtoolbox [Brainard, 1997]. The listeners were asked the following question: *Is the voice disguised or not?*. They responded by clicking “Yes” or “No” on a computer screen. Table-3.5 shows the percentage of subjects who rated each voice as non-disguised. The voices for all three impersonators are included together, so subscripts are used to indicate the impersonator (right-hand numeral) and voice number (left-hand numeral). \( V_{32} \), for example, refers to the third voice of impersonator 2M. The voices are arranged
by their objective rating scores (described below), which are presented in parallel. Overall, the table confirms the prediction that the natural voices (in bold) would receive very high scores. $V_{11}$ and $V_{12}$ are tied with three other voices for the highest score (94.4%), while somewhat surprisingly, $V_{13}$ was judged to be natural only 77.8% of the time. For the impersonated voices, listeners correctly judged these as disguised only 56% of the time, which is only somewhat better than chance. It is important to note that impersonators were not specifically instructed to avoid disguise detection, thus this test provides an estimate of the lower bound on the ability of impersonators to deceive human listeners.

### 3.2.3.2 Objective evaluation

The objective metric of voice disguise that we present here is motivated by the results of Section 3.2.2.3 which showed that there exists an interaction between voice and vowel category. Specifically, the impersonators made changes to the first two formant frequencies on a vowel-by-vowel basis, suggesting that any successful automated system needs to be sensitive to the linguistic parse. Our results suggested that the vowels associated with the impersonators’ natural voices tended to exhibit less variability than the artificial (or disguised) voices, thus we speculated that higher variability might be an important feature associated with voice disguise. This idea is supported by a number of findings in the linguistics literature showing that variability in speech production is closely tied to routinization and practice [Edwards et al., 2004, Munson, 2001]. In [German et al., 2013], this prediction is generated directly from general facts about the organization of the phonological grammar. In short, the impersonators are less practiced with the vowel patterns associated with their artificial voices, and should therefore exhibit more variability in those patterns. We therefore developed an objective metric based on the first two vowel formants that assesses variability across the vowels associated with a voice, but in a way that does not depend on the same type of variability occurring in different vowels (e.g., a systematic increase in $F1$, or a systematic movement towards or away from the center of the vowel space). To accomplish this, the metric makes reference to the linguistic (phonemic) parse, and in doing so remains robust to the voice-by-vowel interaction observed in our earlier findings.

The metric we propose is based directly on the distribution of vowels for a single voice in the $F1$-$F2$ plane and is calculated as follows:

Let $F$ be a matrix which contains the formant values associated with a voice for a vowel $v$. It is defined as

$$F = \begin{bmatrix} f_{1,1} & f_{2,1} \\ f_{1,2} & f_{2,2} \\ \vdots & \vdots \\ f_{1,n} & f_{2,n} \end{bmatrix}$$  \hspace{1cm} (3.6)
Table 3.5: Subjective and Objective ratings of the voices of the three impersonators.

<table>
<thead>
<tr>
<th>Voice</th>
<th>$\zeta$</th>
<th>Subjective Rating (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{63}$</td>
<td>0.8551</td>
<td>83.3</td>
</tr>
<tr>
<td>$V_{11}$</td>
<td><strong>0.8215</strong></td>
<td><strong>94.4</strong></td>
</tr>
<tr>
<td>$V_{12}$</td>
<td><strong>0.8204</strong></td>
<td><strong>94.4</strong></td>
</tr>
<tr>
<td>$V_{32}$</td>
<td>0.8122</td>
<td>94.4</td>
</tr>
<tr>
<td>$V_{13}$</td>
<td><strong>0.8101</strong></td>
<td><strong>77.8</strong></td>
</tr>
<tr>
<td>$V_{81}$</td>
<td>0.8040</td>
<td>72.2</td>
</tr>
<tr>
<td>$V_{52}$</td>
<td>0.7991</td>
<td>16.7</td>
</tr>
<tr>
<td>$V_{22}$</td>
<td>0.7889</td>
<td>33.3</td>
</tr>
<tr>
<td>$V_{23}$</td>
<td>0.7887</td>
<td>38.9</td>
</tr>
<tr>
<td>$V_{43}$</td>
<td>0.7868</td>
<td>61.1</td>
</tr>
<tr>
<td>$V_{92}$</td>
<td>0.7778</td>
<td>94.4</td>
</tr>
<tr>
<td>$V_{62}$</td>
<td>0.7740</td>
<td>5.6</td>
</tr>
<tr>
<td>$V_{41}$</td>
<td>0.7685</td>
<td>66.7</td>
</tr>
<tr>
<td>$V_{83}$</td>
<td>0.7648</td>
<td>66.7</td>
</tr>
<tr>
<td>$V_{91}$</td>
<td>0.7631</td>
<td>94.4</td>
</tr>
<tr>
<td>$V_{31}$</td>
<td>0.7601</td>
<td>22.2</td>
</tr>
<tr>
<td>$V_{33}$</td>
<td>0.7523</td>
<td>0.0</td>
</tr>
<tr>
<td>$V_{53}$</td>
<td>0.7451</td>
<td>22.2</td>
</tr>
<tr>
<td>$V_{61}$</td>
<td>0.7449</td>
<td>44.4</td>
</tr>
<tr>
<td>$V_{52}$</td>
<td>0.7349</td>
<td>5.6</td>
</tr>
<tr>
<td>$V_{71}$</td>
<td>0.7294</td>
<td>27.8</td>
</tr>
<tr>
<td>$V_{51}$</td>
<td>0.7283</td>
<td>22.2</td>
</tr>
<tr>
<td>$V_{21}$</td>
<td>0.7253</td>
<td>22.2</td>
</tr>
<tr>
<td>$V_{72}$</td>
<td>0.7237</td>
<td>16.7</td>
</tr>
<tr>
<td>$V_{42}$</td>
<td>0.6945</td>
<td>5.6</td>
</tr>
<tr>
<td>$V_{03}$</td>
<td>0.6908</td>
<td>55.6</td>
</tr>
<tr>
<td>$V_{73}$</td>
<td>0.6793</td>
<td>27.8</td>
</tr>
</tbody>
</table>
where the columns of \( F \) represent the F1 and F2 values in hertz respectively and the rows represent tokens (samples). Each column of \( F \) is zero mean centered. Then by applying Principal Component Analysis (PCA) we find the reconstruction of \( F \) in the principal component space denoted as \( \hat{F} \). Let \( \sigma_1 \) and \( \sigma_2 \) denote the standard deviation along the two columns of \( \hat{F} \), where \( \sigma_1 > \sigma_2 \). The deviation factor \( \alpha \) for the vowel \( v \) is then defined as

\[
\alpha_v = \frac{\sigma_1}{\sigma_1 + \sigma_2}
\]

(3.7)

where \( 0.5 < \alpha_v < 1 \).

To obtain the overall disguise score for a voice, the deviation factor \( \alpha_v \) is first obtained for \( N \) vowels. The disguise score \( \zeta \) for a voice is then defined as the average value of \( \alpha_v \) over the \( N \) vowels

\[
\zeta = \frac{\sum_{i=1}^{N} \alpha_{v_i}}{N}
\]

(3.8)

Our analysis uses the same set of vowels from Section 3.2.2.3, namely, \( v = \{/æ/, /\alpha/, /\lambda/, /\alpha/, /\#/, /\#/, /\#/\} \).

The disguise score \( \zeta \) provides an estimate of how much a voice varies along the first principal component compared to the total variation along both the principal components. A value of \( \zeta \) near 1.0 is predicted for undisguised (i.e., natural) voices, while a value of \( \zeta \) near 0.5 is predicted for poorly disguised voices. Table-3.5 shows the objective rating score for each voice ordered by the value of \( \zeta \), alongside the subjective ratings from the previous section. As predicted, the natural voices (in bold) are highly ranked. Interestingly, the natural voice for 3M is ranked more highly by \( \zeta \) than by the human raters, suggesting a possible human bias to which it is immune. The table also suggests a good correlation between the objective metric and the subjective ratings. The Spearman rank order correlation coefficient [Spearman, 1904] was calculated to determine the relationship between \( \zeta \) and the subjective ratings of disguise. This statistic measures the strength of a monotonic relationship between two ordinal variables. This test revealed a “strong”, positive correlation between \( \zeta \) and the subjective ratings, which was also statistically significant (\( r_s(25) = 0.6542, p = 8.0144 \times 10^{-4} \)).

Overall, these results indicate that \( \zeta \) is a good indicator of whether a voice is disguised or not. This metric makes use of vowel formants in combination with a linguistic parse, two factors which are generally ignored by automated speaker recognition systems, but which are essential for phoneticians in their manual analyses. In doing so, it provides an assessment of disguise that is highly comparable to that of human listeners, and may even outperform them in certain cases.
3.3 Conclusion

In this chapter, we analysed the speech of three voice impersonators producing a total of 27 different voice identities. These analyses confirmed that the impersonators were able to exploit differences in mean F0, speech rate, vocal fold patterns (Open Quotient), and vowel formant distributions in order to create the various voice identities. This is the first study we know of that shows the ability of voice impersonators to exploit differences in vocal fold patterns, and the only study that considers a comprehensive set of speech parameters in a single study based on voice impersonation. Our study also sought to explore the space of variability that is possible for the various speech parameters given the impersonators’ sensitivity to naturalness constraints and their inherent physiological limitations. On the one hand, by eliciting a wide range of voice identities from the impersonators’ repertoire, our study provides an estimate of the size of the range that is possible for each parameter for a given speaker. This range is not continuous but provides the limits within which a speaker makes modifications to his natural voice. Additionally, our study showed speaker-by-speaker limitations, since, for example, 1F was unable to achieve F0 means comparable to the two male speakers even when impersonating males.

Our analysis of the effects of voice identity on vowel formant measures revealed that while impersonators exploit variation in vowel formants, they do so in a way that is sensitive to linguistic structure. Specifically, they make changes to formant distributions on a vowel-by-vowel basis, rather than by systematically shifting the entire vowel space along some dimension, or by expanding or contracting the vowel space. We noted that this has important consequences for automated disguise detection systems, since it suggests that such systems cannot do without a linguistic parse. This is the first study we know of to show that the modification of vowel formants by impersonators is sensitive to, and constrained by, the specific structure of the linguistic system (i.e., the language) involved.

Auditorily, the voice impersonators were highly successful in changing their voice identity (as confirmed by the subjective test) in terms of age, gender and voice quality, and our study showed that these changes were accompanied by changes in mean F0, intonation, speaking rate and the vocal tract shape to achieve different voice identities. On the one hand, our findings highlight the overall instability of these parameters relative to a particular speaker, and thereby underscore the scope of the challenge presented to automated speaker recognition.

We also presented an objective metric for detecting voice disguise. This metric not only rated the impersonators’ natural voices very highly, but it exhibited a strong correspondence with the subjective ratings obtained from human listeners (even outperforming them in one instance). A fully automatic voice detection system would require a voice disguise feature vector together with a classification system. In the next chapter, we present the mathematical formulation of an auto-
matic voice detection system. Motivated by the results in this chapter, this system utilizes a feature vector related to the vowel variances with the incorporation of a statistical classification method.
Detecting Voice Disguise from Speech Variability

The deliberate attempt by speakers to conceal their identity (voice disguise) presents a challenge for forensics and for automated speaker identification systems. Using a database of natural and disguised voices of three professional voice artists, we build on earlier findings in Chapter 3 where the acoustic variability related to vowel formants was found to be a useful metric of voice naturalness. Here we investigate whether the vowel formants can be useful way for automatic discrimination between natural and disguised voices of impersonators. We also investigate whether there are any features from these measures that can be useful for discriminating natural and disguised voices. By modeling the overall variability of speakers in the vowel space, we propose a new method for automatically classifying natural and disguised voices of speakers. The proposed method is found to outperform several state-of-the-art methods.

This chapter is organized as follows: in Section 4.1 the vowel space analysis is presented. In Section 4.2, we first describe the baseline method for disguise detection and then present a new method, the results are discussed in Section 4.4 and finally Section 4.5 concludes the chapter.

4.1 Vowel space analysis

In this section, we present an analysis of the vowel space. The vowel space depends on the variations in the vocal tract shape and is useful for understanding the strategies and variability exhibited by the artists when producing different vowels. Here we demonstrate how the variability of voices in the vowel space can be a
useful parameter for separating natural and disguised voices. Acoustic variability such as that related to formant variances in vowels has been shown to be an important factor for speech recognition and intelligibility [Wade et al., 2007]. Based on some of our earlier findings in Chapter 3, we predict that the artists will be more practised (less variable) in their natural voice vowel productions than their disguised voices.

The vowel space is a two dimensional plot of the first two formant frequencies (F1 and F2). F1 and F2 are the primary acoustic correlates of perceptual differences among vowel categories. A vowel space is therefore useful for visualizing the inter- and intra-speaker variability exhibited by speakers in their vowel productions. If we consider the disguised and natural voices of the artists as two separate groups in the vowel space, then a greater inter-class separation and smaller intra-class variability is desirable for accurate discrimination between the disguised and natural voices.

The vowel space analysis was performed by first obtaining the first two formant frequencies (F1 and F2) for the target vowels. These were estimated using the Burg method in Praat [Boersma and Weenink, Retrieved October 21, 2011].

Figure 4.1: The natural and disguised voices for vowel /u/. The ellipses represent the 95% confidence region.
With the number of poles set to 12, the F1 and F2 values were obtained from the temporal center of each target vowel. A frequency window of 0-5.5 kHz was used with an analysis window length of 25 ms. The F1 and F2 values obtained were then manually checked, with less than 7% values identified as having potentially erroneous formant estimates based on what is typical for each vowel. These values were then corrected by a visual inspection of the Short Time Fourier Transform (STFT) time-frequency distribution (spectrogram).

The F2 vs. F1 plot of vowel /u/ for the three natural voices and a single disguised voice of each artist is shown in Figure 4.1. The ellipses in Figure 4.1 represent the 95% confidence region with natural and disguised voices represented by red and blue colored ellipses, respectively. The vowel tokens have not been shown here for the sake of clarity, however, readers can refer to Figure 3.13, for example, which shows the ellipses and vowel tokens for all voices of 1F. From Figure 4.1, we find that the natural and disguised voices of the artists occupy similar regions in the F1-F2 space. The artist 1F in particular shows significant overlap between her natural and disguised voice. Considerable overlap between the disguised and natural voices was also observed for the rest of the target vowels as well. From these results, it is clear that any approach which relies on the formant values or in general the spectral features will have a very high chance of confusing disguised and natural voices. Another key observation from Figure 4.1 is that the ellipses associated with natural voices occupy much smaller area in the vowel space compared to the disguised voice ellipses. This suggests that the artists were more consistent in their natural voices when producing the target vowels. A multivariate analysis of variance (MANOVA) also confirmed that the variances in F1 and F2 associated with natural voices were among the lowest when averaged across all vowels. Moreover, there was no evidence found suggesting that the artists were simply changing their vocal tract length for producing different voices. In Chapter 3, it was also found that the artists made adjustments to the formats on a vowel-by-vowel basis, which is unlikely to be explained by a simple contraction or expansion of the vowel space. In Section 4.2.2, we build upon these findings and show how the vowel space variability can be modeled as a feature vector for discriminating disguised and natural voices.

4.2 Methods

In this section, we first describe the state-of-the-art baseline method which has already been used in [Perrot and Chollet, 2008] for the detection of disguised and natural voices. Then we propose a new method which relies on the vowel space variability for discriminating natural and disguised voices.
4.2. METHODS

4.2.1 MFCC-GMM method

The most popular method used for speaker identification [Reynolds and Rose, 1995] and speaker verification [Reynolds et al., 2000] involves use of Mel Frequency Cepstral Coefficients (MFCCs) [Davis and Mermelstein, 1980] as features together with a Gaussian Mixture Model (GMM) as a supervised classifier. The MFCC-GMM approach was also used in [Perrot and Chollet, 2008] for the purpose of automatic disguise detection.

For an \( L \) dimensional feature vector \( x \) the Gaussian mixture density is a weighted sum of \( M \) unimodal Gaussian densities and is defined as

\[
p(x|\lambda) = \sum_{k=1}^{M} w_k p_k(x).
\]  

(4.1)

Here \( w_k \) are the weights and \( p_k \) represents the individual Gaussian densities. The sum of all the weights must be equal to one, i.e. \( \sum_{k=1}^{M} w_k = 1 \). The Gaussian density function is defined as

\[
p_k(x) = \frac{1}{(2\pi)^{L/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu_k)' (\Sigma_k)^{-1} (x - \mu_k) \right\},
\]  

(4.2)

where \( \mu_k \) is a \( L \times 1 \) vector of means and \( \Sigma_k \) is a \( L \times L \) covariance matrix. Thus, a Gaussian mixture model can be completely described by a set \( \lambda \) as

\[
\lambda = \{p_k, \mu_k, \Sigma_k\}, \quad \text{where } k = 1, \ldots, M.
\]  

(4.3)

Given a set of feature vectors, the expectation-maximization algorithm [Dempster et al., 1977] can then be used to obtain the maximum likelihood estimates of the model parameters given by \( \lambda \).

For the purpose of disguise detection, two Gaussian mixture models are constructed: one for the natural voices denoted as \( \lambda_N \) and one for the disguised voices denoted as \( \lambda_D \). To perform the binary classification of a sequence of feature vectors \( X = \{x_1, x_2, \ldots, x_T\} \), the following likelihood ratio test is used

\[
\Lambda(X) = \log p(X|\lambda_N) - \log p(X|\lambda_D).
\]  

(4.4)

Thus if \( \Lambda(X) \geq 0 \) the feature vectors \( X \) belong to a natural voice and if \( \Lambda(X) < 0 \) the feature vectors \( X \) belong to a disguised voice. For a sequence of feature vectors \( X \), the log likelihood function in Equation (4.4) can be calculated as the summation of the logarithm of the Gaussian density function in Equation (4.1). It is thus defined as

\[
\log p(X|\lambda) = \sum_{t=1}^{T} \log p(x|\lambda).
\]  

(4.5)
4.2. METHODS

The GMM classifier is typically used together with MFCC feature vectors which are computed over short-time windows over the voiced segments of the speech signal. The MFCCs provide an efficient representation of the spectral envelope and are considered free from the influences of the vocal folds. The dimensionality \( L \) of the MFCCs is usually 36 including the delta and double-delta coefficients [Reynolds and Rose, 1995].

4.2.2 Ellipse-QDA method

Now we show how the vowel variances in the F1-F2 space can be modeled for the purpose of disguise detection. Let \( G^k \) be the matrix which contains the first two formants for a vowel \( k \). It is defined as

\[
G^k = \left[ f_1^k \ f_2^k \right]
\]  

(4.6)

where \( f_1^k \) and \( f_2^k \) are \( n \times 1 \) vectors representing the F1 and F2 values of vowel \( k \) in Hertz, respectively,

\[
f_1^k = \left[ f_{11}^k f_{12}^k \ldots f_{1n}^k \right]^\top
\]  

(4.7)

and

\[
f_2^k = \left[ f_{21}^k f_{22}^k \ldots f_{2n}^k \right]^\top
\]  

(4.8)

where \( n \) are the number of tokens (samples) of vowel \( k \).

As a first step, we make the vectors \( f_1^k \) and \( f_2^k \) zero mean

\[
f_1^k = f_1^k - \bar{f}_1^k
\]  

(4.9)

\[
f_2^k = f_2^k - \bar{f}_2^k
\]  

(4.10)

The matrix \( G^k \) in Equation (4.6) represents a set of points (tokens) of a vowel \( k \) in the F1-F2 space. The arrangement, position and variability of the vowel \( k \) can be captured by drawing an ellipse around these points. The ellipse can be constructed such that it covers, for example, 95\% of the data points as shown in Figure 4.2.

Now we show how the relevant parameters of such an ellipse can be extracted given the matrix \( G^k \). Let \( C^k \) be the correlation matrix for vowel \( k \) defined as

\[
C^k = \begin{bmatrix}
\langle f_1^k, f_1^k \rangle & \langle f_1^k, f_2^k \rangle \\
\langle f_2^k, f_1^k \rangle & \langle f_2^k, f_2^k \rangle
\end{bmatrix}.
\]  

(4.11)

Let \( \nu_1^k \) and \( \nu_2^k \) represent the eigen vectors and \( \alpha_1^k \) and \( \alpha_2^k \) represent the eigen values of the correlation matrix in Equation (4.11), where \( \alpha_1^k \geq \alpha_2^k \). The eigen
4.2. METHODS

Figure 4.2: The 95% confidence ellipse for vowel /æ/. The major and minor axis of the ellipse are represented by $\gamma_1$ and $\gamma_2$.

Vectors here represent the axes of the ellipse. The coordinates of the points projected on these axes follow a Gaussian distribution with zero mean and variance which is given as

$$\sigma_1^k \equiv \sqrt{\frac{\alpha_1^k}{n-1}} \quad (4.12)$$

$$\sigma_2^k \equiv \sqrt{\frac{\alpha_2^k}{n-1}}. \quad (4.13)$$

Let $U^k$ and $V^k$ represent the coordinates of the ellipse which lie along the direction of $v_1^k$ and $v_2^k$ respectively. Then the ellipse is of the form

$$Z^k = \left(\frac{U^k}{\sigma_1^k}\right)^2 + \left(\frac{V^k}{\sigma_2^k}\right)^2. \quad (4.14)$$

The ellipse in Equation (4.14) follows a chi-squared distribution. For a 95% confidence ellipse, using a lookup table we find that $P(Z^k < 5.991) = 0.95$. Using this in Equation (4.14), we get

$$\left(\frac{U^k}{\sigma_1^k}\right)^2 + \left(\frac{V^k}{\sigma_2^k}\right)^2 = 5.991. \quad (4.15)$$
From Equation (4.15), we find that the major and minor axis of the ellipse which covers 95% of the data points are given by

\[
\gamma_k^1 = \sqrt{5.991} \sigma_k^1
\]

(4.16)

\[
\gamma_k^2 = \sqrt{5.991} \sigma_k^2.
\]

(4.17)

Using these major and minor axis of the 95% ellipse, the feature vector \( \delta^k \) for the \( k^{th} \) vowel is defined as

\[
\delta^k = [\gamma_k^1 \gamma_k^2].
\]

(4.18)

Given a set of \( N \) vowels, the feature vector for a voice is then defined as

\[
\beta = \sum_{k=1}^{N} \delta_k.
\]

(4.19)

Here \( \beta \) represents the overall variation exhibited by a voice in the vowel space. Ellipses having smaller area in the vowel space will have smaller values of \( \delta_k \) in both dimensions. Smaller values of \( \beta \) on the other hand imply an "overall" smaller area occupied by a voice in the vowel space. Thus natural voices are predicted to have smaller values \( \beta \) (higher consistency) compared to disguised voices. Given the lower dimensionality of \( \beta \) compared to the MFCC features (2 vs. 36), the Quadratic Discriminant Analysis (QDA) was found suitable for classification purposes. The QDA classifier uses a quadratic boundary to separate different classes [Hastie et al., 2009]. It also has the advantage of having a lower computational efficiency than the GMM classifier.

### 4.3 Experimental setup

The speech data used for all the experiments is the same as described in Section 3.2.1. For the MFCC-GMM method, the MFCC feature vectors were computed over all the voiced segments of the speech signals. The number of Gaussians, \( M \), was set to 32. Increasing the number of Gaussians did not result in an increase in either the disguise or natural voice detection accuracy. For the proposed Ellipse-QDA method, the number of vowels, \( N \), was set to 6. The vowels used were: /æ/, /ɛ/, /i/, /ɪ/, /u/, and /ɛ/. The results are presented next.

### 4.4 Results

In this section, we compare the performance of the methods which were presented in Section 4.2, that is, the MFCC-GMM method and our newly proposed ellipse based disguise detection method. In order to have a more fair and thorough comparison, we also compared the proposed method with two other variants of the
Table 4.1: The confusions matrices for the the different matrices indicating their performance for disguised and natural voices. A high performing system has higher values along the diagonal entries and lower values of the off diagonal entries.

<table>
<thead>
<tr>
<th>Natural</th>
<th>Disguise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>100 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>4.17 %</td>
</tr>
<tr>
<td>Natural</td>
<td>58.28 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>55.17 %</td>
</tr>
</tbody>
</table>

(a) Ellipse-QDA

<table>
<thead>
<tr>
<th>Natural</th>
<th>Disguise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>7.59 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>5.43 %</td>
</tr>
<tr>
<td>Natural</td>
<td>15.40 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>14.76 %</td>
</tr>
</tbody>
</table>

(c) MFCC-QDA

<table>
<thead>
<tr>
<th>Natural</th>
<th>Disguise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>100 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>95.83 %</td>
</tr>
<tr>
<td>Natural</td>
<td>58.28 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>44.83 %</td>
</tr>
</tbody>
</table>

(b) MFCC-GMM

<table>
<thead>
<tr>
<th>Natural</th>
<th>Disguise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>7.59 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>94.56 %</td>
</tr>
<tr>
<td>Natural</td>
<td>15.40 %</td>
</tr>
<tr>
<td>Disguise</td>
<td>85.23 %</td>
</tr>
</tbody>
</table>

(d) MFCC-SVM

MFCC based method, which use a Quadrature Discriminant Analysis classifier (MFCC-QDA) and a Support Vector Machine classifier (MFCC-SVM). This is to ensure that the gain in performance observed for the proposed method is not due to a difference in the type of method used for classification. Table-4.1 shows the performance of all the different methods in terms of confusion matrices. Larger values of the diagonal entries and smaller values of the off diagonal elements are desirable for a high performing system. From Table-4.1, we find that the Ellipse-QDA method correctly classifies a high percentage (100 and 95.83) of natural and disguised voices while only confusing a small number of disguised voices (4.17 %) as being natural. The MFCC-GMM system on the other hand is found to be at best guessing about the voice type while both the MFCC-QDA and MFCC-SVM systems seem to be overfitting the disguised voices. Thus their accuracy for correctly detecting natural voices is very low (7.59 % and 15.40 % respectively). Together, the results indicate a) the superiority of using the vowel variability features over the spectral features and b) the usefulness of using only key segments of speech (target vowels) rather than utilizing all voiced portions.

4.5 Conclusions

We proposed a new approach based on utilizing the variability of the vowel distributions of speakers in the formant space. Compared to traditional approaches which model the spectral envelope for the voiced segments of speech, we instead model the overall variability of a voice for specific vowels. The results indicate that for the task of discriminating natural and disguised voices, the proposed vowel
consistency features clearly outperform the traditional spectral features. In conclusion, vowel variability is found to be highly important feature for the discrimination of disguised and natural voices.
In Chapter 4, the problem of automatic disguise detection was presented. However, disguised voices also present a challenge in the context of speaker recognition i.e., how can we reliably predict the identity of the speaker in situations where the speaker is deliberately disguising his/her voice to avoid detection. In Chapter 3, we demonstrated how different vowels are subject to different kinds of variations in the vowel space and how the impersonators made adjustments to vowel formants on a vowel by vowel basis. Inspired from this, in this chapter, we focus on developing speaker recognition models which model different units of speech (phones) separately, so-called the phonetic categories approaches for speaker recognition.

The baseline method based on Gaussian Mixture Modeling of speakers is first presented in Section 5.1, followed by two existing phonetic models of speaker recognition in Section 5.2. We then introduce a new model in Section 5.3 which overcomes some of the drawbacks of the existing models. The discrimination capabilities of the different phones are discussed in Section 5.5. The experimental conditions are discussed in Section 5.4 and the performance of the various speaker recognition models is discussed in Section 5.6. Finally Section 5.7 concludes the chapter.
5.1 Single GMM based speaker recognition

The baseline approach for speaker recognition or identification [Reynolds and Rose, 1995] uses MFCCs as features and then models each speaker using a Gaussian Mixture Model (GMM). We refer to this baseline method as the Single GMM (SGMM) speaker recognition system. This model is described by the set of parameters $\lambda$ as defined in Equation (4.3).

Suppose there are $N$ speakers and $T$ training samples $\{x_t, y_t\}, t = 1, \ldots, T$ where $x_t \in \mathbb{R}^L$ and $y_t \in \mathbb{R}^1$. Here $x_t$ is a $L$ dimensional feature vector and $y_t$ is the corresponding speaker label. The expectation-maximization algorithm [Dempster et al., 1977] can then be used to obtain the maximum likelihood estimates of the speaker model parameters given by $\lambda$. Thus, after training, each speaker $i$ in the system is represented by a single GMM model $\lambda_i$. In the testing phase, when a test feature vector $x_v$ is presented to the system, identification is performed by finding the speaker model $\lambda_i$ which maximizes the posterior probability. The speaker $i$ belonging to $\lambda_i$ is thus the identified speaker. Mathematically, this can be written as

$$y_v = \arg \max_{1 \leq i \leq N} \Pr (\lambda_i | x_v) = \arg \max_{1 \leq i \leq N} \frac{p(x_v | \lambda_i) \Pr (\lambda_i)}{p(x_v)}. \quad (5.1)$$

Assuming equally likely speakers Equation (5.1) simplifies to

$$y_v = \arg \max_{1 \leq i \leq N} p(x_v | \lambda_i) \quad (5.2)$$

5.2 Phonetic GMM speaker recognition

The SGMM model works well if there is broad and rich phonetic coverage in the training data of speakers, however, this is often not the case especially in forensic scenarios where there is often a shortage of data. Although the Gaussian densities in a SGMM model are somewhat representative of the entire phonetic space, the SGMM model does not explicitly make use of any phonetic information. As a result, having only one GMM model per speaker fails to capture the acoustic variability of phonemes in the training data when comparing different speakers [Faltlhauser and Ruske, 2001, Park and Hazen, 2002]. Also, several studies have already reported that different phonemes exhibit different speaker separating capabilities. In [Eatock and Mason, 1994, Faltlhauser and Ruske, 2001], for example, it was found that nasals /n/ and /m/ provided higher identification rates while having relatively lower frequency of occurrence than other phonemes, thereby indicating their superiority over other phonemes. Because of these reasons, a Phonetic GMM (PGMM) speaker recognition system represents each speaker with multiple GMM models instead of only one as in the SGMM model. The motiva-
tion is to capture fine grained speaker specific phonetic information which might be lost in a SGMM model.

Given a set of $P$ phonetic classes and $N$ speakers, training is performed such that each speaker $i$ is now represented by $P$ GMM models defined as $\lambda_i^k$ for $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, P$.

Depending on how the PGMM system is tested, two different kinds of PGMM models can be constructed namely the Phonetic GMM-Matched model and the Phonetic GMM-Unmatched model. The details of these two models are presented next.

5.2.1 Phonetic GMM-Matched model

In a Phonetic GMM-Matched (PGMM-M) model, the phone label of the test feature vector is known in advance. This allows a test feature vector belonging to a particular phonetic class to be only compared with the corresponding GMM phone models for that phone across all speakers [Park and Hazen, 2002]. The speaker whose GMM phone model that gives the maximum likelihood score is predicted to be the recognized speaker. Figure 5.1 shows the architecture of the PGMM-M model.

The PGMM-M speaker recognition model can also be described mathematically. Let $x_{Cv}$ represent a test feature vector belonging to phonetic class $C$. The feature vector is compared against the corresponding phone model $\lambda_i^C$ to obtain the scores $S_{Cv}^i$ given as

$$S_{Cv}^i = \log p \left( x_{Cv} \mid \lambda_i^C \right), \quad i = 1, 2, \ldots, N.$$  (5.3)

The logarithm in Equation 5.3 is used to make the distribution of the scores normal. The recognized speaker $y_{Cv}$ corresponding to the feature vector $x_{Cv}$ is then given as

$$y_{Cv}^C = \arg \max_{1 \leq i \leq N} S_{Cv}^i.$$  (5.4)

5.2.2 Phonetic GMM-Unmatched model

In contrast to the PGMM-M model, the Phonetic GMM-Unmatched (PGMM-U) model does not require the phone label of the test feature vector to be known in advance. Thus in the PGMM-U model, a test feature (from any phone class) is compared with all the GMM phone models across all speakers. The speaker is then identified by taking the maximum or the sum of the phone likelihoods [Faltlhauser and Ruske, 2001]. Figure 5.2 shows the architecture of the PGMM-M model.
Figure 5.1: Flow diagram of the Phonetic GMM-Matched model. The input frame is only compared to its corresponding phone model.
Figure 5.2: Flow diagram of the Phonetic GMM-Unmatched model. The input frame is compared to all phone model.
5.3. PHONETIC GMM-EXTREME LEARNING MACHINE METHOD

The PGMM-U speaker recognition model can also be described mathematically. Let $x_v$ represent a test feature vector and $S_{i,v}$ represent the score of each speaker for $x_v$. The score $S_{i,v}$ can be obtained by either taking the sum or the maximum of the speaker phone likelihoods. Thus, it can be defined as either

$$S_{i,v} = \sum_{k=1}^{P} \log p \left( x_v | \lambda^k_i \right), \quad i = 1, 2, \ldots, N \quad (5.5)$$

or in case of taking the maximum of the speaker phone likelihoods, as

$$S_{i,v} = \max \log p \left( x_v | \lambda^k_i \right), \quad i = 1, 2, \ldots, N. \quad (5.6)$$

The recognized speaker $y_v$ is then given as the one having the maximum score

$$y_v = \arg \max_{1 \leq i \leq N} S_{i,v}. \quad (5.7)$$

5.3 Phonetics GMM-Extreme Learning Machine method

The two phonetic models for speaker recognition introduced earlier in Section 5.2 have their own advantages and disadvantages. The PGMM-M model requires a phonetic label of the test feature vector and thus allows for direct comparisons of the test feature vector with the corresponding speaker GMM phone models. Because of this, the PGMM-M model in principle should provide more accurate estimates of speaker identity than the PGMM-U model. However, this comes at the cost of requiring an accurate phonetic parse of the test feature vectors which is still an active research area and still requires further improvement [Kim and Conkie, 2002, Toledano et al., 2003]. The PGMM-U model, on the other hand, does not require a phonetic labeling of the feature vectors during testing. As a result, it requires comparisons of the input feature vector with all the GMM phone models across all speakers which leads to a decrease in the system accuracy.

In the Phonetic GMM-Extreme Learning Machine method, we aim to achieve higher speaker recognition accuracy without the need of requiring a phonetic segmentation of the test feature vectors. Notice that for the PGMM-U model, the sum or maximum of the phone likelihoods (Equation (5.5), Equation (5.6)) is used for the computation of the speaker scores. Thus, the PGMM-U model does not make use of all the available information for the computation of the speaker scores. Retaining all the available information can be useful for cases where a simple sum or maximum of the phone likelihoods does not always give the highest score for the correct speaker. For such cases, it might be that some other complex non-linear relationship of the phone-likelihoods exists. Thus to overcome these limitations
5.3. PHONETIC GMM-EXTREME LEARNING MACHINE

METHOD

of the PGMM-U model, we propose using Extreme Learning Machine (ELM) classifier paired with the the PGMM-U model for speaker recognition, so-called the PGMM-ELM model. The PGMM-ELM model retains the phone likelihoods across all speakers and then utilizes the ELM classifier to find a non-linear function best discriminating the speakers. An overview of the ELM classifier is presented next followed by a description of the PGMM-ELM model.

5.3.1 Extreme Learning Machine classifier

The Extreme Learning Machine (ELM) [Huang et al., 2006] is a supervised classifier inspired from single-hidden layer feedforward networks (SLFNs). Compared to SLFNs, the training time of ELM is very small. It also offers advantages such as avoiding local minima, improper learning rate and overfitting. Thus, it offers good generalization at very high learning speeds. The basic concept of SFLN is presented next which is followed by a description of the ELM.

Suppose there are \(N\) number of classes (speakers) and \(T\) training samples \(\{x_t, y_t\}, t = 1, \ldots, T\) where \(x_t \in \mathbb{R}^L\) and \(y_t \in \mathbb{R}^N\). Here \(x_t\) is a \(L\) dimensional feature vector and \(y_t\) is the corresponding coded speaker class label. The objective of training a classifier then is to learn a mapping function \(y = F(x)\) which maps the input feature vectors to their corresponding target labels. For a SLFN with \(q\) hidden neurons, the approximation \(\hat{y}\) of the mapping function is defined as follows

\[
\hat{F}(x_t) = \sum_{i=1}^{q} \beta_i G(\alpha_i, b_i, x_t)
\]  

(5.8)

where \(G(\alpha_i, b_i, x)\) represents the output of the \(i\)th hidden neuron. \(\alpha_i\) are the \(L \times q\) weights from the input layer to the hidden layer, \(\beta_i\) represents the \(q \times N\) output weights and \(b_i\) are the \(q \times 1\) biases of hidden neurons. Figure 5.3 shows the architecture of SFLN.

In case of sigmoidal hidden neurons with activation function \(g(x)\), the output of the \(i\)th hidden neuron is given as

\[
G(\alpha_i, b_i, x) = g(\alpha_i \cdot x_t + b_i).
\]  

(5.9)

For Radial Basis Function (RBF) hidden neurons with activation function \(g(x)\), the output of the \(i\)th hidden neuron is given as

\[
G(\alpha_i, b_i, x) = g(b_i \|x_t - \alpha_i\|)
\]  

(5.10)

where \(\alpha_i\) and \(b_i\) (\(b_i \in \mathbb{R}^+\)) represent the center and impact factor of the \(i\)th RBF hidden neuron, respectively.

For a group of \(T\) training samples, Equation (5.8) can be rewritten as

\[
\sum_{i=1}^{q} \beta_i G(\alpha_i, b_i, x_t) = y_t, \quad t = 1, \ldots, T.
\]  

(5.11)
These $T$ set of equations can be rewritten in matrix form as

$$H\beta = Y$$  \hspace{1cm} (5.12)

where

$$H = \begin{bmatrix} G(\alpha_1, b_1, x_1) & \cdots & G(\alpha_q, b_q, x_1) \\ \vdots & \ddots & \vdots \\ G(\alpha_1, b_1, x_T) & \cdots & G(\alpha_q, b_q, x_T) \end{bmatrix}$$  \hspace{1cm} (5.13)

and

$$\beta = \begin{bmatrix} \beta_1^Y \\ \vdots \\ \beta_q^Y \end{bmatrix}, \quad Y = \begin{bmatrix} y_1^Y \\ \vdots \\ y_T^Y \end{bmatrix}.$$  \hspace{1cm} (5.14)
5.3. PHONETIC GMM-EXTREME LEARNING MACHINE

METHOD

Here $H$ is the $T \times q$ hidden layer output matrix, $\beta$ is the $z \times N$ weights matrix from the hidden layer to the output and $Y$ is the $T \times N$ matrix of the target labels.

The number of hidden neurons are generally much smaller than the number of training samples, i.e., $q \ll T$. In the ELM algorithm [Huang et al., 2006], the weights $\beta$ and biases $b$ are selected randomly. The SLFN is thus trained by finding a least square solution $\hat{\beta}$ of the linear system in Equation (5.12). It is given as

$$\hat{\beta} = H^\dagger Y$$  

(5.15)

where $H^\dagger$ is the Moore-Penrose inverse [Albert and Albert, 1972] of $H$.

The ELM algorithm can be summarized as follows:

1. The values for $\alpha_i$ and $b_i$ are randomly assigned.
2. Compute the hidden layer output matrix $H$.
3. Calculate the output weights $\hat{\beta}$ using Equation (5.15).

5.3.2 The PGMM-ELM model

The PGMM-ELM model is an extension of the PGMM-U model. For a text feature vector, the PGMM-ELM model retains the GMM phone likelihoods across all speakers, fuses these together to form a new feature vector which is then used as the input to the ELM classifier. Figure 5.4 shows the architecture of the PGMM-ELM model.

The PGMM-ELM model can also be described mathematically. Given an input feature vector $x_v$, the score for a speaker $i$ is defined as

$$S_{i,v} = \left[ \log p(x_v|\lambda_{1}^i) \quad \log p(x_v|\lambda_{2}^i) \quad \ldots \quad \log p(x_v|\lambda_{P}^i) \right]$$  

(5.16)

for $i = 1, 2, \ldots, N$.

The scores from all the speakers are then combined to form a new feature vector

$$z_v = [S_{1,v} \quad S_{2,v} \quad \ldots \quad S_{N,v}]$$  

(5.17)

where $z_v$ is an $NP$ dimensional feature vector for the $L$ dimensional input feature vector $x_v$. This new feature vector thus contains the phone likelihood information for all the phones across all speakers. The ELM is then trained using a set of $T$ training samples $(z_t, y_t)$ for $t = 1, 2, \ldots, T$. After the ELM has been trained, a test feature vector $z_v$ is presented to the classifier which then returns the identified speaker as $y_v$ according to the procedure described in Section 5.3.1.

Next, we introduce two variants of the basic PGMM-ELM model, namely the PGMM-Weighted ELM (PGGM-WELM) model and the PGMM-Reduced ELM (PGGM-RELM) model.
Figure 5.4: The architecture of the WELM speaker recognition model.
5.3.3 The PGMM-WELM model

One motivation of using the PGMM based speaker recognition models is to make use of the fact that different phones can offer different amounts of speaker discriminating capabilities. Thus, a natural way would be to weight different phones according to their performance based on some criterion. The PGMM-WELM model is similar to the PGMM-ELM model except that it applies a weighting to the phone likelihoods in Equation (5.17). The phones which are found to have higher speaker discriminating accuracy or rate (obtained from prior knowledge) are weighted more highly compared to phones with lesser speaker discriminating accuracy. Thus, for the PGMM-WLEM model, the feature vector $z^W_v$ is defined as

$$z^W_v = \begin{bmatrix} S_{1,v} \cdot W \quad S_{2,v} \cdot W \quad \ldots \quad S_{N,v} \cdot W \end{bmatrix}. \quad (5.18)$$

Here $W$ is a $P$ dimensional phone weight vector

$$W = \begin{bmatrix} w_1 & w_2 & \ldots & w_P \end{bmatrix}. \quad (5.19)$$

The weights are normalized, such that, $\sum_{k=1}^{P} w_k = 1$. The same set of weights $W$ are applied for each speaker and can be obtained, for example, by using the phone recognition rates obtained from the PGMM-U model. These are given in Table 5.2. Other strategies to find weights include minimization of a cost function described in [Faltlhauser and Ruske, 2001]. More details on the phone weighting will be presented in Section 5.5.

5.3.4 The PGMM-RELM model

The size of the feature vector $z_v$ introduced in Section 5.3.2 depends on the number of speakers $N$ and the number of phone classes $P$. As $N$ and/or $P$ increase, the size of $z_v$ increases. In the PGMM-RELM model, we aim to improve the computational efficiency of the system by reducing the number of dimensions of the feature vector from $NP$ to $N + P$. Let $z^R_v$ denote the feature vector for the PGMM-RELM model. We first reorganize the $NP$ dimensional vector $z^R_v$ into a $N$ by $P$ matrix $U_v$, such as

$$U_v = \begin{bmatrix} \log p(x_v|\lambda_1^1) & \log p(x_v|\lambda_1^2) & \ldots & \log p(x_v|\lambda_1^P) \\ \log p(x_v|\lambda_2^1) & \log p(x_v|\lambda_2^2) & \ldots & \log p(x_v|\lambda_2^P) \\ \vdots & \vdots & \ddots & \vdots \\ \log p(x_v|\lambda_N^1) & \log p(x_v|\lambda_N^2) & \ldots & \log p(x_v|\lambda_N^P) \end{bmatrix}. \quad (5.20)$$

Now let $\bar{U}_v$ be a vector which is composed of the $P$ mean values obtained from $U_v$. It is defined as

$$\bar{U}_v = \begin{bmatrix} \mu_v^1 & \mu_v^2 & \ldots & \mu_v^P \end{bmatrix}. \quad (5.21)$$
where
\[
\mu_k^v = \sum_{i=1}^{N} \log p(x_v | \lambda_i^k) / N \quad k = 1, 2, \ldots, P. \tag{5.22}
\]

The mean vector $\bar{U}_v$ represents the average phone likelihood variation over all speakers. The feature vector $z^R_v$ is then given as
\[
z^R_v = [S_{1,v}, S_{2,v}, \ldots, S_{N,v}, \bar{U}_v] \tag{5.23}
\]
where $S_{i,v}$ is given by Equation (5.5) or Equation (5.6). The $N + P$ vector $z^R_v$ is thus composed of two parts, the first $N$ values are the speaker scores while the next $P$ values represent the average of the phone likelihoods. As a result, if the number of speakers in the system increase, the growth in the size of $z^R_v$ will be much smaller than $z_v$. Another strategy to reduce the size of ELM feature vector can be to reduce the number of phones $P$ by grouping different phones into one GMM model based on some optimization criterion.

5.4 Experimental setup

The speech data used in the experiments included a) the sentences described in Section 3.2.1 and b) an approximately 1 minute long paragraph given in Appendix A.

For all experiments, the Mel Frequency Cepstral Coefficients (MFCCs) were used as features. A 25 ms frame size was set and a Hamming window was used to obtain 12 MFCCs for each frame. There was a 50% overlap between consecutive frames. In total, 36 MFCCs were used including delta and delta-delta coefficients.

Diagonal covariance matrices were used for all GMMs. The number of Gaussians were set to 256 for the SGMM model while 8 GMMs per phone model were used for the PGMM-M, PGMM-U and the PGMM-ELM models. Increasing the number of Gaussians beyond these values did not show any improvement in performance. For the PGMM-U model the maximum of the scores as given is Equation (5.6) was always used as it was found to perform better than using the sum of the phone likelihoods. For the PGMM-ELM based models the input feature vectors during testing and training were normalized before being fed to the ELM classifier. The number of hidden neurons $q$ was set to 300 and the sigmoidal activation function was used. For all speaker recognition models, only the natural voices of the impersonators were used for training when testing for disguised voices.

5.5 Discriminative capabilities of phones

One of the main motivations of using the PGMM speaker recognition models is to make use of the fact that different phones can contribute differently to the
Table 5.1: The phones used during analysis and their time duration.

<table>
<thead>
<tr>
<th>IPA symbol</th>
<th>Time duration (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>13.4046</td>
</tr>
<tr>
<td>a</td>
<td>13.0488</td>
</tr>
<tr>
<td>i</td>
<td>9.0075</td>
</tr>
<tr>
<td>d</td>
<td>8.1762</td>
</tr>
<tr>
<td>ñ</td>
<td>7.6404</td>
</tr>
<tr>
<td>n</td>
<td>6.8942</td>
</tr>
<tr>
<td>k</td>
<td>6.3925</td>
</tr>
<tr>
<td>Æ</td>
<td>6.0642</td>
</tr>
<tr>
<td>t</td>
<td>5.8400</td>
</tr>
<tr>
<td>a</td>
<td>5.2696</td>
</tr>
<tr>
<td>u</td>
<td>4.9542</td>
</tr>
<tr>
<td>z</td>
<td>4.6121</td>
</tr>
<tr>
<td>ë</td>
<td>4.3596</td>
</tr>
<tr>
<td>o</td>
<td>4.1183</td>
</tr>
<tr>
<td>r</td>
<td>3.9729</td>
</tr>
<tr>
<td>ai</td>
<td>3.9104</td>
</tr>
<tr>
<td>e</td>
<td>3.8425</td>
</tr>
<tr>
<td>w</td>
<td>3.8008</td>
</tr>
<tr>
<td>et</td>
<td>3.6213</td>
</tr>
<tr>
<td>b</td>
<td>3.3500</td>
</tr>
<tr>
<td>ao</td>
<td>3.2842</td>
</tr>
<tr>
<td>ø</td>
<td>3.2725</td>
</tr>
<tr>
<td>p</td>
<td>3.2679</td>
</tr>
<tr>
<td>f</td>
<td>2.9471</td>
</tr>
<tr>
<td>m</td>
<td>2.9125</td>
</tr>
<tr>
<td>h</td>
<td>2.2496</td>
</tr>
<tr>
<td>η</td>
<td>2.1717</td>
</tr>
<tr>
<td>ŋf</td>
<td>1.9708</td>
</tr>
<tr>
<td>oʊ</td>
<td>1.7446</td>
</tr>
<tr>
<td>ð</td>
<td>1.7000</td>
</tr>
<tr>
<td>v</td>
<td>1.6117</td>
</tr>
<tr>
<td>ðɣ</td>
<td>1.5500</td>
</tr>
<tr>
<td>ñ</td>
<td>1.4596</td>
</tr>
</tbody>
</table>
Table 5.2: The recognition rates (RR) of the phones in percentage. The results are shown here for both the natural and disguised voices of impersonators obtained from the PGMM-M and PGMM-U models.

<table>
<thead>
<tr>
<th>Phone</th>
<th>RR</th>
<th>Phone</th>
<th>RR</th>
<th>Phone</th>
<th>RR</th>
<th>Phone</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td></td>
<td>Disguise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PGMM-M</td>
<td></td>
<td>PGMM-U</td>
<td></td>
<td></td>
<td>PGMM-M</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>9.4433</td>
<td></td>
<td>20.7914</td>
<td></td>
<td></td>
<td>9.3193</td>
<td></td>
</tr>
<tr>
<td>λ</td>
<td>7.8129</td>
<td></td>
<td>9.1085</td>
<td></td>
<td></td>
<td>8.1694</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>6.1508</td>
<td></td>
<td>5.8976</td>
<td></td>
<td></td>
<td>5.8976</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>5.1292</td>
<td></td>
<td>5.8828</td>
<td></td>
<td></td>
<td>5.5795</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>4.9998</td>
<td></td>
<td>3.9252</td>
<td></td>
<td></td>
<td>4.8286</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>4.3626</td>
<td></td>
<td>3.8475</td>
<td></td>
<td></td>
<td>3.9995</td>
<td></td>
</tr>
<tr>
<td>Æ</td>
<td>3.5812</td>
<td></td>
<td>2.8724</td>
<td></td>
<td></td>
<td>3.2771</td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>3.3479</td>
<td></td>
<td>2.5239</td>
<td></td>
<td></td>
<td>3.3854</td>
<td></td>
</tr>
<tr>
<td>æ</td>
<td>3.3203</td>
<td></td>
<td>2.0716</td>
<td></td>
<td></td>
<td>2.7096</td>
<td></td>
</tr>
<tr>
<td>æ</td>
<td>3.1896</td>
<td></td>
<td>2.0608</td>
<td></td>
<td></td>
<td>2.6909</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.6397</td>
<td></td>
<td>1.9706</td>
<td></td>
<td></td>
<td>2.6774</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.4817</td>
<td></td>
<td>1.7903</td>
<td></td>
<td></td>
<td>2.6612</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.4555</td>
<td></td>
<td>1.7728</td>
<td></td>
<td></td>
<td>2.4747</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.3355</td>
<td></td>
<td>1.7626</td>
<td></td>
<td></td>
<td>2.4560</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.2751</td>
<td></td>
<td>1.6767</td>
<td></td>
<td></td>
<td>2.2718</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.1965</td>
<td></td>
<td>1.4927</td>
<td></td>
<td></td>
<td>2.1868</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.1469</td>
<td></td>
<td>1.4764</td>
<td></td>
<td></td>
<td>2.0009</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>2.1112</td>
<td></td>
<td>1.4732</td>
<td></td>
<td></td>
<td>1.8456</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.9645</td>
<td></td>
<td>1.4682</td>
<td></td>
<td></td>
<td>1.6565</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.8155</td>
<td></td>
<td>1.4574</td>
<td></td>
<td></td>
<td>1.6292</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.6676</td>
<td></td>
<td>1.4146</td>
<td></td>
<td></td>
<td>1.6292</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.5005</td>
<td></td>
<td>1.3927</td>
<td></td>
<td></td>
<td>1.4323</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.4951</td>
<td></td>
<td>1.0173</td>
<td></td>
<td></td>
<td>1.2864</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.1387</td>
<td></td>
<td>1.0047</td>
<td></td>
<td></td>
<td>1.0799</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.0839</td>
<td></td>
<td>0.8820</td>
<td></td>
<td></td>
<td>0.9332</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>1.0352</td>
<td></td>
<td>0.8378</td>
<td></td>
<td></td>
<td>0.9192</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>0.9933</td>
<td></td>
<td>0.7962</td>
<td></td>
<td></td>
<td>0.8895</td>
<td></td>
</tr>
<tr>
<td>Ë</td>
<td>0.9076</td>
<td></td>
<td>0.5713</td>
<td></td>
<td></td>
<td>0.8411</td>
<td></td>
</tr>
</tbody>
</table>
accuracy of a speaker recognition system. In other words, some phones will be more robust to attempts of voice disguise as speakers find it hard to manipulate these phones compared to others. Thus, it is useful to do an analysis to understand how different phones contribute to the accuracy of a PGMM speaker recognition system. Such an analysis is useful, for example, to find the phone weights or to combine different phones together into a single GMM model to improve the system performance. In this section, a comparison of the recognition accuracy of different phones for the PGMM-M and PGMM-U models is presented. Table 5.1 shows the phones used for performing the experiments together with their time duration in sorted order. The phones in Table 5.1 were selected based on the criteria of having a duration of at least 1 second. Thus, a total of 33 phones were selected in this way.

Table 5.2 shows the recognition rate (RR) of all the phones for both the PGMM-M and PGMM-U systems. The phones have been ordered in descending order of the RR for the PGMM-M and PGMM-U speaker recognition systems. The RR is normalized and is given in percentage. The RR for a phone is obtained by counting the number of feature vectors for which the phone model gives correct speaker (highest score) and dividing it by the total number of feature vectors for all phones. The RR found in this way is then averaged across all speakers. Thus the RR here indicates how much on average each phone contributes to the overall recognition accuracy. From Table 5.2 we also find that although each phone has some contribution to the accuracy, Table 5.2 shows that some phones contribute significantly higher than others. From Table 5.2 we find that /s, ʌ, t, ɪ, d / are some of the highly rated phones by both systems. From Table 5.1, we find that these phones also have some of the highest overall time durations. However, some phones, such as, /s, r/ for the PGMM-U model display a higher RR (5.8976% and 6.1073%) for the natural and disguised voices respectively while having a relatively smaller duration (4.3596 s and 3.9729 s). Similarly the nasal /n/ also displays a higher RR with a smaller duration for the PGMM-M model for the disguised and natural voices. These results indicate that although some of the higher duration phones achieve higher RR, each phone contributes differently to the other phones. Another important point here is that the phone rankings for each model stay relatively constant across the natural and disguised voices. This means a weighting function can be learned based on the RR of the natural voices which can then be applied when recognizing speakers from their disguised voices. There also exists some correlation between the phone rankings of the PGMM-M and PGMM-U models.

The rankings in Table 5.2 somewhat resemble the ranking found. For the task of speaker recognition, the authors in [Faltlhauser and Ruske, 2001] analyzed only 8 phones and found /m, n, s/ as the top three phones having higher recognition rates (RR). In [Eatock and Mason, 1994], /ŋ, a, i, u/ were having higher RR. These rankings somewhat resemble the ranking found in Table 5.2.
5.6 Comparison of the performance of different speaker recognition models

In this section, we compare the performance of the different speaker recognition models which have been described earlier. As with any classification system, the performance of a speaker recognition system can be assessed by the Receiver Operating Characteristics (ROC) curve. The ROC curve plots the true positive rate vs. the false positive rate for different thresholds of the classifier outputs. Curves which are nearer the top left hand corner are indicative of a high performing classifier while those which follow the diagonal indicate a classifier doing no better than making random guesses. Figure 5.5 shows the ROC plot for the natural voices of impersonators. Here we find the the PGMM-ELM, PGMM-WELM and PGMM-RELM speaker recognition models perform considerably better than the other models. From Figure 5.5, we also find that the PGMM-M model has better performance than the SGMM and PGMM-U models. This is expected as the PGMM-M model makes direct comparisons of the input feature vector to its corresponding phone model.

For the case of disguised voices, we expect the classification performance to be affected which is exactly what we observe from the ROC plot in Figure 5.6. Here we find that all the speaker recognition models are performing worse than for natural voices in Figure 5.5. However, again the PGMM-ELM, PGMM-WELM and PGMM-RELM models still show a better performance than the rest of the models.

Based on the ROC curve, the performance of the speaker recognition models can also be expressed in terms of the Area Under Curve (AUC). An ideal classifier has an AUC of 1.0 while an AUC of 0.5 indicates a classifier making random guesses. Table 5.4 shows the AUC for the natural and disguised voices for the different speaker recognition models. From Table 5.4, we find that the PGMM-WELM model has the highest AUC for the natural voices. For disguised voices, the AUC decreases for all models. The PGMM-ELM model has the highest AUC value for the disguised voices.

Table 5.3 shows the testing accuracy for the natural and disguised voices for the different speaker recognition models. From Table 5.3, we find that PGMM-WELM has the highest accuracy for recognizing natural voices of speakers while the PGMM-ELM has the highest accuracy for the disguised voices. Again, its clear from Table 5.3 that the the performance of all methods is affected when detecting disguised voices.
5.6. COMPARISON OF THE PERFORMANCE OF DIFFERENT SPEAKER RECOGNITION MODELS

Figure 5.5: Receiver Operating Characteristics plot for the detection of natural voices.

Table 5.3: Testing accuracy in percentage for the natural and disguised voices.

<table>
<thead>
<tr>
<th>Method</th>
<th>Natural</th>
<th>Disguise</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGMM</td>
<td>75.49</td>
<td>67.07</td>
</tr>
<tr>
<td>PGMM-M</td>
<td>76.29</td>
<td>67.29</td>
</tr>
<tr>
<td>PGMM-U</td>
<td>75.84</td>
<td>67.01</td>
</tr>
<tr>
<td>PGMM-ELM</td>
<td>78.44</td>
<td>71.35</td>
</tr>
<tr>
<td>PGMM-WELM</td>
<td>78.56</td>
<td>70.97</td>
</tr>
<tr>
<td>PGMM-RELM</td>
<td>78.49</td>
<td>71.28</td>
</tr>
</tbody>
</table>
5.6. COMPARISON OF THE PERFORMANCE OF DIFFERENT SPEAKER RECOGNITION MODELS

Figure 5.6: Receiver Operating Characteristics plot for the detection of disguised voices.
5.7 Conclusions

In this chapter, several phonetic speaker recognition models were presented. We introduced the PGMM-ELM model which fuses the likelihood information of phones from all speakers to form a feature vector. This feature vector is then fed to the ELM classifier which then outputs a speaker label. We also introduced two variants of the PGMM-ELM model: the PGMM-WELM model which applies weights to the feature vector and the PGMM-RELM model which reduces the dimensionality of the feature vector. The results indicate the fact that disguised voices of speakers have a significant impact on the performance of speaker recognition systems. The proposed method was found to outperform the existing methods when recognizing impersonators from both their natural and disguised voices. Together, the results indicate the usefulness of taking into account phonetic information for speaker modeling as well as for disguise detection as discussed in Chapter 4.

Table 5.4: Area under curve for the natural and disguised voices.

<table>
<thead>
<tr>
<th>Method</th>
<th>Natural</th>
<th>Disguise</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGMM</td>
<td>0.7142</td>
<td>0.6016</td>
</tr>
<tr>
<td>PGMM-M</td>
<td>0.7253</td>
<td>0.6132</td>
</tr>
<tr>
<td>PGMM-U</td>
<td>0.6997</td>
<td>0.5966</td>
</tr>
<tr>
<td>PGMM-ELM</td>
<td>0.8192</td>
<td>0.6735</td>
</tr>
<tr>
<td>PGMM-WELM</td>
<td>0.8246</td>
<td>0.6715</td>
</tr>
<tr>
<td>PGMM-RELM</td>
<td>0.8226</td>
<td>0.6700</td>
</tr>
</tbody>
</table>
Glottal Activity Detection using Finite Rate of Innovation Methods

Electroglottography was first introduced by Fabre in 1957 [Fabre, 1957]. It is a non invasive technique which estimates the vocal fold contact area by measuring the electrical conductance between the vocal folds. Two electrodes are placed on the neck of the subject and a small high frequency electrical current is passed. The output from the electrodes is then fed to the electroglottograph. The signal thus obtained is called an ElectroGlottoGraph (EGG) signal. For voiced phonation, the vocal folds vibrate in a periodic manner, moving in and out of contact with each other. Thus, the EGG signal also varies periodically as a function of the contact area between the vocal folds. However, during unvoiced phonation, the vocal folds stay apart (open glottis) and therefore the EGG signal is a low amplitude, non-periodic signal. The EGG signal thus provides useful information for understanding and analyzing the complex periodic movements of the vocal folds during voiced phonation [Childers and Krishnamurthy, 1985].

Now consider a vocal fold vibratory cycle in which the vocal folds are initially not in contact, thus, the electrical conductance as measured by the electroglottograph is minimum. As the vocal folds begin to move in contact, the electrical conductance starts to increase. The time instant at which the glottis (the space between vocal folds) becomes closed is called the Glottal Closing Instant (GCI). The glottal closing is generally abrupt and appears as a steep slope in the EGG signal as shown in Figure 3.10. It is widely accepted that the GCI appears as a sharp positive peak in the Differentiated Electroglottograph (DEGG) signal [Childers and Krishnamurthy, 1985, Baken, 1992, Cranen, 1988, Henrich et al., 2004]. The glottis then remains closed for a short period of time before the vocal folds start separating again causing the measured electrical conductance to decrease. The
time instant at which the glottis becomes opened is called the Glottal Opening Instant (GOI). The GOI appears as a low amplitude peak in the DEGG with a polarity opposite to that of the GCI peak [Henrich et al., 2004]. The EGG and DEGG signals corresponding to a voiced segment of speech together with the labeled GCIs and GOIs are shown in Figure 3.10. Thus the DEGG signal contains distinct landmarks which can be useful for detection of GCIs and GOIs [Henrich et al., 2004, Howard, 1995].

The process of detecting the GCIs and GOIs is referred to as glottal activity detection. It is useful for many areas of speech processing such as the detection of pathological voices [Ritchings et al., 2002], speaker identification [Campbell et al., 2003], closed phase analysis of speech [Krishnamurthy and Childers, 1986] and voice quality assessment [Fourcin, 2000]. The GCIs and GOIs estimated accurately from the DEGG signal can also serve as the benchmarks for algorithms which estimate them directly from the speech signal [Drugman et al., 2012]. The simplest methods for the detection of GCIs and GOIs use a threshold based approach [Howard, 1995, Huckvale, 2012, Bouzid and Ellouze, 2009], however, they fail to perform well when the DEGG signal is corrupted by noise. This noise caused by electrical interferences during phonation results in the DEGG signals having multiple GCIs and GOIS over a single glottal period [Henrich et al., 2004]. In [Henrich et al., 2004] the DEGG periods with multiple GCI peaks were removed from the analysis and it was found that the number of usable DEGG cycles can be as low as 8%. However, this method of elimination, significantly reduces the amount of useful DEGG cycles. To overcome these issues, we present a method that is robust to noise and provides accurate estimates of GCIs and GOIs while having a very limited number of false alarms.

Accurate detection of GCIs and GOIs is useful for many applications, such as, providing a reference for pitch detection algorithms [Kawahara et al., 1999], speech dereverberation [Gaubitch et al., 2004], voice source modeling [Thomas et al., 2009] and prosodic speech modification [Valbret et al., 1992] etc. It is also useful for the analysis of pathological speech, which includes types of dysphonia [Casper and Leonard, 2006], vocal fold impact stress [Verdolini et al., 1998] and essential tremor [Gamboa et al., 1998]. In Chapter 3, Section 3.2.2.2, the GCIs and GOIs were also used to study the vocal fold characteristics of the impersonators.

This chapter is organized as follows: Section 6.1 presents a Finite Rate of Innovation (FRI) model to estimate the GCIs and GOIs; Section 6.2 describes a scheme for estimating the GCIs and GOIs from the DEGG signals; the performance of the proposed scheme is discussed in Section 6.3; and finally Section 6.4 concludes the chapter.
6.1 Modeling the Glottal Closing and Opening Instants as Finite Rate of Innovation Signals

In this section, we describe an FRI model for the recovery of GCIs and GOIs from a DEGG signal. Let $x_1(t)$ and $x_2(t)$ represent the signals corresponding to the GCI and GOI of a DEGG signal, respectively,

$$x_1(t) = c_{GCI} \delta(t - t_{GCI})$$  \hspace{1cm} (6.1)$$

$$x_2(t) = c_{GOI} \delta(t - t_{GOI})$$  \hspace{1cm} (6.2)$$

where $\{t_{GCI}, t_{GOI}\}$ represent time instants and $\{c_{GCI}, c_{GOI}\}$ represent the amplitudes corresponding to the GCI and GOI. The signals in Equations (6.1), (6.2) together with a DEGG signal are shown in Figure 6.1. As it can be seen from Figure 6.1, the signals $x_1(t)$ and $x_2(t)$ correspond only to the points of interest, i.e. the GCI and the GOI whereas the other details present in the DEGG signal are ignored.

6.1.1 Recovering the GCI and GOI from a DEGG signal

Both $x_1(t)$ and $x_2(t)$ are FRI signals each having two innovations: $\{t_{GCI}, c_{GCI}\}$ and $\{t_{GOI}, c_{GOI}\}$ respectively. In order to obtain the GCIs and GOIs, we need to recover the innovations of the signals $x_1(t)$ and $x_2(t)$ from the samples of a real DEGG signal. Since the signals in Equations (6.1), (6.2) are both non-bandlimited (because of the presence of Diracs), the FRI methods in [Blu et al., 2008] are found appropriate to recover these innovations. Now we explain the process of recovering GCIs given a DEGG signal. We assume that a particular segment (window) of a real DEGG signal is a low pass filtered version of the FRI signal in Equation (6.1). A sinc sampling kernel $\varphi(t)$ with a bandwidth $B$ is used to filter the signal in Equation (6.1). The filtered signal is then uniformly sampled with a sampling period $T = \tau/N$, where $N$ is the number of samples and $\tau$ represents the time period of one cycle of a DEGG signal corresponding to a GCI segment. The sampled signal is then given as

$$y_n = \langle x_1(t), \text{sinc}(B(nT - t)) \rangle = c_{GCI} \varphi(nT - t_{GCI}), \hspace{1cm} n = 1, ..., N$$  \hspace{1cm} (6.3)$$

The Annihilating Filter method in [Blu et al., 2008] is then used to obtain the innovations: $\{t_{GCI}, c_{GCI}\}$ from the $N$ samples in Equation (6.3). This method consists of the following steps:

1. Obtain the $N$-point DFT from the samples given in Equation (6.3)

$$\hat{y}_m = \sum_{n=1}^{N} y_n e^{-j2\pi nm/N}$$  \hspace{1cm} (6.4)$$
6.1. MODELING THE GLOTTAL CLOSING AND OPENING INSTANTS AS
FINITE RATE OF INNOVATION SIGNALS

Figure 6.1: A period of aDEGG signal and the FRI signals $x_1(t)$ and $x_2(t)$. The signals $x_1(t)$ and $x_2(t)$ only capture the points of interest, i.e. the GCI and GOI.
6.2 THE PROPOSED SCHEME

2. Build a rectangular Toeplitz matrix from the DFT coefficients in Equation (6.4).

\[
A = \begin{bmatrix}
\hat{y}_N & \hat{y}_{N-1} \\
\hat{y}_{N-1} & \hat{y}_{N-2} \\
\vdots & \ddots \\
\hat{y}_2 & \hat{y}_1 \\
\end{bmatrix}
\]  

(6.5)

3. Given \(A\), solve for \(Ah = 0\), where \(h\) is the Annihilating filter. The method in [Maravic and Vetterli, 2005] is used to obtain the denoised roots \(e^{-j2\pi t_{GCI}/\tau}\), from which we obtain \(t_{GCI}\).

4. The final step is to compute the least mean square solution \(c_{GCI}\) of the \(N\) equations \(\{y_n - c_{GCI}\phi(nT - t_{GCI})\}_{n=1,2,...,N}\).

Thus, it is possible to recover the GCIs from the samples of a real DEGG signal. A similar process can be applied for the recovery of GOIs. An advantage of the FRI method is that the resolution of the GCIs and GOIs obtained does not depend on the sampling period. This is due to the fact that the \(\{t_{GCI}, t_{GOI}\}\) obtained are not necessarily equal to the sampling locations \(nT\) for any \(n = 1, ..., N\).

The methods presented in this section require that carefully selected segments of a DEGG signal which actually correspond to a GCI or a GOI must be available before processing. However, DEGG signals corresponding to long segments of speech consist of many such segments distributed over voiced portions of the DEGG signal. In the next section, we will present a complete scheme, which carefully selects these segments from the voiced portions before applying the methods presented in Section 6.1 to recover the GCIs and GOIs.

6.2 The proposed scheme

In this section, a complete scheme for the detection of the GCIs and GOIs from long duration DEGG signals containing both voiced and unvoiced intervals is presented. Figure 6.2 shows a flow chart illustrating the various steps involved. The first step consists of separating the voiced and unvoiced segments present in the DEGG signal. For each voiced segmented identified, the GCI and GOI windows are found. The final step is to use FRI methods described in Section 6.1.1 to recover the GCIs and GOIs. The details of the each step involved in GCI/GOI detection are presented next.

6.2.1 Voiced/unvoiced (V/UV) detection

A real DEGG signal consists of intervals corresponding to both voiced and unvoiced components of speech. However, the GCIs and GOIs are distributed only
6.2. THE PROPOSED SCHEME

6.2.2 Detecting the GCIs

Once all the voiced and unvoiced segments present in the DEGG signal have been identified, the next step is to find GCI and GOI windows which will be used to recover the time instants: \( \{t_{GCI}, t_{GOI}\} \) and amplitudes: \( \{c_{GCI}, c_{GOI}\} \) corresponding to the GCIs and GOIs.

6.2.2.1 Finding the GCI windows

The next step, given a voiced segment of the DEGG signal, is to find GCI windows which contain approximately one GCI. A peak detection algorithm is used for this purpose. A point is considered a peak if it has the maximal value and was preceded...
Figure 6.3: Voiced/unvoiced detection for a DEGG signal corresponding to a sentence. The values of ‘V’ and ‘UV’ on the vertical right hand axis indicate if a DEGG segment is voiced or unvoiced respectively.
to the left by a value lower than a threshold $\epsilon$. The threshold $\epsilon$ is set by adding the standard deviation of the unvoiced segments to the standard deviation of the current voiced segment. Thus the threshold for the detection of GCI windows is set adaptively based on SNR of the DEGG signal. This ensures that spurious noisy peaks will not be selected as candidates for further processing. The peak detection algorithm thus provides the window locations $w_{GCI,i}$ where $i = 1, \ldots, L$, and $L$ denotes the total number of window locations detected. The $i$th GCI window is then defined as $[w_{GCI,i} - 1/f_{\text{max}}, w_{GCI,i} + 1/f_{\text{max}}]$, where $f_{\text{max}}$ represents the maximum frequency of voicing.

### 6.2.2.2 Obtaining the final estimates for the GCIs

Once the window locations $w_{GCI,i}$ have been determined, the FRI methods described in Section 6.1 are then used to recover the $t_{GCI,i}$ and $c_{GCI,i}$ corresponding to each window. Often the EGG signal gets corrupted by noise due to swallowing or subject movement or due to electrical interferences of the equipment used during the recording process. This noise can result in the DEGG signal having spurious Dirac like impulses which might be wrongly identified as the GCIs and GOIs. Therefore, a post-processing step is performed to remove them. For each $t_{GCI,i}$, if $1/f_{\text{max}} < t_{GCI,i} - t_{GCI,i-1} < 1/f_{\text{min}}$ and $1/f_{\text{max}} < t_{GCI,i+1} - t_{GCI,i} < 1/f_{\text{min}}$, keep the $t_{GCI,i}$, else remove it.

### 6.2.3 Detecting the GOIs

For the estimation of GOIs, the location of the GOI windows, i.e. $w_{GOI,i}$, needs to be determined. The $t_{GCI,i}$ obtained in Section 6.2.2.2 are used for this purpose. The $i$th GOI window is defined as $[t_{GCI,i} + 0.5/f_{\text{max}}, t_{GCI,i+1} - 0.5/f_{\text{max}}]$. Once the windows $w_{GOI,i}$ have been determined the same process as in Section 6.2.2.2 is repeated to obtain the $t_{GOI,i}$ and $c_{GOI,i}$ for each window.

### 6.3 Results and discussion

To evaluate the performance of the proposed algorithm, the APLAWD database [Lindsey et al., 1987] was used. It consists of synchronous speech and EGG recordings of five short sentences spoken ten times by five male and five female speakers. The hand-labeled GCIs and GOIs for this database served as the ground truth. The performance measures are the same as in [Naylor et al., 2007] and include hit, miss, false alarm, false alarm total, hit accuracy and hit bias.

For each period corresponding to a voiced segment of the DEGG signal, we expect only one GCI/GOI. Thus, for each period, a hit occurs when only GCI/GOI is detected, a miss occurs when no GCI/GOI is detected and a false alarm (FA) occurs when more than one GCI/GOI is detected. The false alarm total (FAT)
Table 6.1: GCI detection performance on the APLAWD database

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>Miss</th>
<th>FA</th>
<th>FAT</th>
<th>Hit acc (μsec)</th>
<th>Hit bias (μsec)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRI</td>
<td>99.30</td>
<td>0.55</td>
<td>0.15</td>
<td>0.42</td>
<td>77.76</td>
<td>−22.37</td>
<td>99.23</td>
</tr>
<tr>
<td>SIGMA</td>
<td>97.78</td>
<td>0.07</td>
<td>2.13</td>
<td>2.71</td>
<td>43.40</td>
<td>−23.24</td>
<td>95.27</td>
</tr>
<tr>
<td>Multi-scale</td>
<td>93.64</td>
<td>6.32</td>
<td>0.02</td>
<td>0.03</td>
<td>60.13</td>
<td>−18.75</td>
<td>93.61</td>
</tr>
<tr>
<td>TXGEN</td>
<td>94.78</td>
<td>3.47</td>
<td>1.76</td>
<td>2.63</td>
<td>450.00</td>
<td>−130.00</td>
<td>93.27</td>
</tr>
</tbody>
</table>

Table 6.2: GOI detection performance on the APLAWD database

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>Miss</th>
<th>FA</th>
<th>FAT</th>
<th>Hit acc (μsec)</th>
<th>Hit bias (μsec)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRI</td>
<td>99.28</td>
<td>0.53</td>
<td>0.17</td>
<td>0.16</td>
<td>433.30</td>
<td>153.63</td>
<td>99.12</td>
</tr>
<tr>
<td>SIGMA</td>
<td>99.21</td>
<td>0.05</td>
<td>0.72</td>
<td>0.73</td>
<td>344.01</td>
<td>−14.31</td>
<td>98.49</td>
</tr>
<tr>
<td>Multi-scale</td>
<td>92.21</td>
<td>7.71</td>
<td>0.06</td>
<td>0.06</td>
<td>337.98</td>
<td>138.48</td>
<td>92.14</td>
</tr>
<tr>
<td>TXGEN</td>
<td>94.63</td>
<td>3.55</td>
<td>1.82</td>
<td>2.05</td>
<td>860.00</td>
<td>−50.00</td>
<td>93.05</td>
</tr>
</tbody>
</table>

represents the number of false alarms which are not hits. Hit accuracy and hit bias are calculated by finding the Root Mean Square Error (RMSE) and mean errors between the estimated GCIs/GOIs and the hand-labeled GCIs/GOIs respectively. Table 6.1 and Table 6.2 show the performance of the proposed algorithm for the detection of GCIs and GOIs respectively on the APLAWD database. Three other methods: SIGMA [Thomas and Naylor, 2009], Multi-scale product method [Bouzid and Ellouze, 2009] and Time of Excitation Generator (TXGEN) [Huckvale, 2012] were compared to the proposed scheme. The values of \( f_{\text{min}} \) and \( f_{\text{max}} \) were set to 50Hz and 400Hz respectively as these limits were found to be suitable for the APLAWD database. From Table 6.1 and Table 6.2, it can be seen that the performance of the proposed scheme offers significant improvement over the other the SIGMA and the Wavelets method achieving an overall performance of 97% and 99% for the GCIs and GOIs respectively while only having a very small percentage of misses and false alarms. The hit accuracy is not as good as the SIGMA. This could be due to a) the particular way the GCIs and GOIs were hand-labeled and b) because the hand-labeled GCI and GOI time instants were selected to fall on sampled time instants in [Thomas and Naylor, 2009]. The proposed method on the other hand can provide GCI and GOI time instants which are not restricted to sampling instants which explains the difference in the hit accuracy performance of SIGMA and the proposed method. Nevertheless the hit accuracy is in the micro seconds range indicating a very high correlation between the hand-labeled and estimated GCI/GOI time instants.
6.4 Conclusion

We proposed a new method for the accurate detection of GCIs/GOIs from the DEGG signals. An FRI model of the GCI and GOI signals has been presented which models them as Diracs. The Annihilating filter method is then used to recover the time instants and amplitudes corresponding to the GCIs and GOIs. The experimental results show that proposed method achieves very high hit rates (about 99%) for GCI/GOI detection while having very small miss rates (0.55% and 0.53%) and false alarm rates (0.15% and 0.17%), when tested on database with a total of 500 EGG signals. This method was also used in Chapter 3 for the analysis of the vocal fold timing characteristics of the impersonators. In particular, the analysis showed that the impersonators use variation in open quotient for producing different voice identities, thus suggesting its usefulness for speaker identity. Accurate methods for the detection of GCIs and GOIs are therefore necessary and important to study the vocal fold characteristics of speakers.
Conclusions and Future Work

7.1 Conclusions

In this thesis, a series of linguistic and acoustic analyses of voice impersonations was presented. We analysed the speech and electroglottograph signals of three voice impersonators producing a total of 27 different voice identities. These analyses confirmed that the impersonators were able to exploit differences in mean F0, F0 variance, speech rate, vocal fold patterns (Open Quotient), and vowel formant distributions in order to create the various voice identities. This is the first study we know of that shows the ability of voice impersonators to exploit differences in vocal fold patterns, and the only study that considers a comprehensive set of speech parameters in a single study based on voice impersonation. The study also sought to explore the space of variability that is possible for the various speech parameters given the impersonators’ sensitivity to naturalness constraints and their inherent physiological limitations. On the one hand, by eliciting a wide range of voice identities from the impersonators’ repertoire, our study provides an estimate of the size of the range that is possible for each parameter for a given speaker. Additionally, our study showed speaker-by-speaker limitations, since, for example, impersonator 1F was unable to achieve F0 means comparable to the two male speakers even when impersonating males.

In terms of the analysis, the relationship between voice identity and vowel characteristics showed complex interdependencies that can only be accounted for by a consideration of linguistic factors. For example, differences in vocal tract length associated with an individual’s body size are predicted to systematically raise or lower all formants in a speaker’s vowel inventory, yet the different voices produced by the impersonator affected the formants in a vowel-dependent way. In some cases, these dependencies appear to reflect the overall arrangement of

99
the vowels over the vowel space, and the fact that the impersonators were relying on their own linguistic knowledge to ensure that the distributions of the different vowels remained distinct for any one voice. In other cases, however, the voice-by-vowel interactions may reflect the voice impersonators own judgements regarding the relative naturalness of different types of formant modifications. This suggests that the problems, such as, speaker recognition, disguise detection etc. require a separate consideration of different voice parameters in terms of their relationship to the linguistic system.

The analyses also showed that separate linguistic and acoustic factors are resources that the voice impersonator exploits to achieve different voice identities. Some factors are truly linguistic (e.g., the difference between post-vocalic /r/ and non-postvocalic /r/), some are purely non-linguistic (like various voice-quality parameters), while some are related to linguistic factors, but gradient and within-category (meaning that the language tolerates some variability within a range, which can be exploited for social reasons without affecting the linguistic parse).

Motivated by the analyses, novel methods for two different biometric applications, namely disguise detection and speaker recognition, were also presented. By incorporating linguistically relevant features (vowels), we showed that the vowel variability such as that related to vowel variances in the F1-F2 space is a good indicator of voice disguise. In other words, substantial variance was manifested by the voice impersonators in the acoustic parameters and the degree of variance particularly that related to F1-F2 plane was used as an objective measure of voice disguise. This metric not only rated the impersonators’ natural voices very highly, but it also exhibited a strong correspondence with the subjective ratings obtained from human listeners. The analysis of the voices in the F1-F2 space also showed that the natural voices of impersonators manifested much less variability in comparison to the disguised voices. We then proposed a new method for automated disguise detection which used the variability related to vowel variances as a feature vector together with a quadratic discriminant analysis classifier. This new method for automated disguise detection was found to outperform existing methods as well as human listeners for the same task.

We also presented several phonetic methods for recognizing impersonators from their disguised and natural voices. In particular, the PGMM-ELM method was introduced which fuses the phone likelihood information from different phones into a feature vector thereby retaining all the linguistic information. The feature vector is then fed to extreme learning machine for classification which is computationally efficient and thus capable of handling high dimensional data. We found the PGMM-ELM method is able to outperform other methods for speaker recognition which do not take sufficient linguistic information into consideration. The proposed method also needed a very short duration of speech (25 ms) to predict the identity of the speaker. The results obtained from this study are extremely useful and apply directly to forensic speaker recognition which is still a develop-
ing field and requires manual interpretation by a trained forensic phonetician. A fully automated forensic speaker recognition system is extremely useful as it is free from observer bias and also computationally efficient.

In conclusion, using acoustic and linguistic information together provides a better understanding of the various glottal, vocal tract and vocal fold characteristics of voice impersonations. Moreover, the combination of the two also improves the classification performance of disguise detection and forensic speaker recognition systems.

### 7.2 Future Work

In our study of disguise detection, we selected a subset of vowels and treated the contribution of all vowel categories equivalently. In Chapter 5, we showed how different phonemes possess different amounts of speaker recognition capabilities. These findings suggest the usefulness of additional studies that begin to explore and model the variability in other units of speech, for example, consonants. For vowels, we utilized the variability in the F1-F2 space, however, for other categories of phonemes this may not be suitable. Therefore, a through study of the acoustic properties of different phoneme categories can be conducted to identify the relevant features in different phonemes for modeling variability. Different portions of the spectral envelope, for example, can be studied in isolation to find the one’s which offer the most discrimination between the disguised and natural voices of speakers. In short, other spaces of variation can be investigated and explored which may further be useful for voice disguise detection. Finally the variability computed from different phonetic categories of speech can be fused together to form a a robust feature vector for disguise detection. Therefore, a method that uses variability in different phonetic categories of speech may provide better resolution and therefore better overall performance.

In addition, the weighting of vowels used in calculating the proposed disguise detection feature $\beta$ can also be investigated and varied to improve performance. The weights can be found such that the vowels which have better accuracy for separating disguised and natural voices are weighted more highly than others. Thus, a method that applies weighting to different vowels depending on their contribution to disguise discrimination may also lead to further gains in reliability. Improvements in performance of the segmentation and formant tracking algorithms will eliminate the reliance on hand-correction, thereby allowing the proposed method to be readily implemented in a fully-automated way.

For the problem of forensic speaker recognition, it will be important in future to compare the results of the proposed method to those from a forensic expert to further justify its usefulness and reliability in criminal cases. Another approach might explore ways to group different phones together instead of modeling them individually. This can be done iteratively by trying different combinations of
phones together and then picking the one which offers the best accuracy. Phones can also be combined by means of optimizing some cost function which minimizes the error rate in recognizing speakers. Combining phones also serves to reduce the dimensionality of the feature vector $z_v$ which improves the computational efficiency of the system.

For the problem of forensic speaker recognition, a perception study can also be conducted by which the ability of human listeners to identify speakers from their disguised voices can be investigated. Listeners can be first trained using the natural voices of impersonators and then presented with their disguised voices for identification. One strategy can be to present listeners with the target words for identification. This can be useful, for example, to study the relationship between the perception of speaker identity under disguise and the various acoustic parameters, such as, F0, vowel formants, open quotient etc. The ultimate goal would be to find a set of perceptually motivated features which are robust to voice disguise and can further result in improvement of forensic speaker recognition systems.

In this study, we focused on using and modeling the segmental information (discreet units of speech) for the purposes of disguise detection and forensic speaker recognition. For disguise detection in particular, we used target vowels to study the acoustic variability of speakers. Having the labels available for different vowel tokens allowed us to study the variability within each vowel category independently. This approach can be also be extended to include suprasegmental information, such as to study the variability in the intonational patterns of speakers. Similar to having vowel labels, the autosegmental metrical theory [Goldsmith, 1990] can be used to provide labels for various intonational patterns. Different intonational target categories can be formed and the variability of the speakers within each category can be studied. The same framework can also be applied to the problem of forensic speaker recognition where the suprasegmental information can also be modeled together with the segmental information. The ultimate system will thus incorporate both segmental and suprasegmental information for various speaker modeling and discrimination applications.
Speech Materials

List of Sentences

1. The little boy’s **dad** had a large collection of **beads**.

2. Mrs. Schumann asked the gardener if he recognized the **buds**, and he swore that he **did**.

3. Whenever Josh is feeling **bad**, he tries to do at least one good **deed**.

4. Though the agent was a bit of a **dud**, John was enthusiastic about the **bid**.

5. There once was a lady named **Sid** who was often mistaken for a **dude**.

6. In spite of the newly planted **seeds**, the garden appeared **dead**.

7. Sasha saw the treasure lying on the **bed**, and it made her feel **sad**.

8. When the jester slipped on the **suds**, the restless crowd **booed**.
9. If Georgia had lost the **deed**, the bank might have **sued**.

**Paragraph**

So far, we have been describing speech sounds by stating how they are made, but it is also possible to describe them in terms of what we can hear. The way in which we hear a sound depends on its acoustic structure. We want to be able to describe the acoustics of speech for many reasons. Linguists and speech pathologists need to understand how certain sounds become confused with one another. We can give better descriptions of some sounds by describing their acoustic structures rather than by describing the articulatory movements involved. A knowledge of acoustic phonetics is also helpful for understanding how computers synthesize speech and how speech recognition works. Furthermore, often the only permanent data that we can get of a speech event is an audio recording, as it is often impossible to obtain movies or x-rays showing what the speaker is doing. Accordingly, if we want permanent data that we can study, it will often have to come from analyzing an audio recording.
List of Author’s publications

Journals


Conferences


**Abstracts**


**Open access layman articles**

Bibliography


