# Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters

Erica Ashley Gold

Submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

The University of York

Department of Language and Linguistic Science

Submitted January 2014

# Abstract

The research presented in this thesis examines the calculation of numerical likelihood ratios using phonetic and linguistic parameters derived from a corpus of recordings of speakers of Southern Standard British English. The research serves as an investigation into the development of the numerical likelihood ratio as a medium for framing forensic speaker comparison conclusions. The thesis begins by investigating which parameters are claimed to be the most useful speaker discriminants according to expert opinion, and in turn examines four of these 'selected/valued' parameters individually in relation to intra- and inter-speaker variation, their capacities as speaker discriminants, and the potential strength of evidence they yield. The four parameters analyzed are articulation rate, fundamental frequency, long-term formant distributions, and the incidence of clicks (velaric ingressive plosives). The final portion of the thesis considers the combination of the four parameters under a numerical likelihood ratio framework in order to provide an overall likelihood ratio.

The contributions of this research are threefold. Firstly, the thesis presents for the first time a comprehensive survey of current forensic speaker comparison practices around the world. Secondly, it expands the phonetic literature by providing acoustic and auditory analysis, as well as population statistics, for four phonetic and linguistic parameters that survey participants have identified as effective speaker discriminants. And thirdly, it contributes to the forensic speech science and likelihood ratios for forensics literature by considering what steps can be taken to conceptually align the area of forensic

speaker comparison with more developed areas of forensic science (e.g. DNA) by creating a human-based (auditory and acoustic-phonetic) forensic speaker comparison system.

# Table of Contents

# List of Tables and Figures

## List of Tables

## List of Figures

# List of Appendices

# Acknowledgements

There is an ancient African proverb that says, “it takes a village to raise a child.” A portion of that proverb – “it takes a village”- is now colloquially used to acknowledge the influence a group of people can have when contributing to something bigger than themselves.

Despite appearing as the single author of this piece of work, I believe that it takes not just a single individual, but a “village” to bring a PhD to fruition. To that extent, I would like to take the opportunity to thank *my* village.

The first thank you goes to my supervisor, Professor Peter French, whose support, guidance, and encouragement has helped shape every bit of this research. Thank you for every discussion (of which there are far too many to even count), the time you have spent editing drafts/abstracts/papers, and all the opportunities you have graciously allowed me.

Thank you is also extended to Professor Paul Foulkes and Dr. Dominic Watt for their advice and suggestions on this thesis. A general thank is also due to the York team for helping to secure the University’s position in the BBfor2 project.

I must thank BBfor2 for being a wonderful learning environment over the past years. An important thank you goes to Dr. David van Leeuwen and Dr. Henk van den Heuvel for their leadership and organization of the project. A thank you is also extended to my supervisor in the BBfor2 network, Dr. Didier Meuwly, for providing valuable insight into the application of Bayes’ Theorem in forensics.

During my PhD I was fortunate enough to go on two placements, from which I learned so much. Thanks go to those who helped me at the Netherlands Forensic Institute, especially Dr. Marjan Sjerps, Jacob de Zoete, and Vikram Doshi with all things Bayesian.

I would also like to thank those at the University of Canterbury in the New Zealand Institute for Language, Brain and Behaviour, especially Professor Jen Hay for allowing me to be a part of such an amazing linguistic environment. Thank you to Dr. Kevin Watson and Dr. Lynn Clark for giving me the opportunity to work alongside you both.

A special thank you goes to Professor Colin Aitken, to whom I sent an email containing a complicated statistics question in early 2012. If I could have anticipated what was to follow, then I probably would have emailed sooner. That email began what was to become an extremely rewarding exchange of ideas at the interface of phonetics and statistics. Thank you for taking the time to reply.

A thank you is also extended to those who have helped me along the way through discussions, insight, and simply encouragement. Thank you to Professor Anders Eriksson, Phil Harrison, Dr. Michael Jessen, Dr. Christin Kirchhübel, Dr. Carmen Llamas, Professor Francis Nolan, Dr. Richard Ogden (for all things non-pulmonic), and Lisa Roberts.

I must also express my gratitude to all 36 anonymous forensic experts who participated in my survey for this PhD research. Without them my PhD would lack a solid basis.

To all my office buddies over the years, I thank you for putting up with my annoying finger tapping (as I counted syllables for AR) and the clicking sounds I made (trying to determine place of articulation), the discussions we have had, and simply from keeping me from talking to myself: Natalie Fecher, Jessica Wormald, Becky Taylor, and Rana Alhussein Almbark.

A thank you must also go to Vincent Hughes for all our LR discussions. Without a fellow LR-researcher this thesis just would not be the same. A thank you also goes to the Forensic Research Group in the Language and Linguistic Science Department at the University of York for providing a safe and nurturing environment in which to exchange ideas.

Now to those behind the scenes. I must thank my family - Mom, Dad, and Lauren – for your unequivocal love and support. You have always encouraged me to follow my interests, so much so that your encouragement has never faltered even with me being 6,000 miles away from home. And sorry for all those middle of the night phone calls when I miscalculated the time difference.

To Tom, my best friend and biggest supporter, you know the details of this PhD almost as well as I do. Thank you for being my shoulder to cry on when things got stressful, lending your ear as I discussed a long day's worth of work, providing unyielding patience, and your continued support and unconditional love.

Finally, despite my love for phonetics and linguistics, the work reported in this thesis would not have been possible without the financial support of the Bayesian Biometrics for Forensics (BBfor2) Network, funded by Marie Curie Actions (EC Grant Agreement No. PITN-GA-2009-238803). Without the generous support of this funding body, this thesis would not be possible.

To my village – I thank you.

# Declaration

This is to certify that this thesis comprises original work and that all contributions from external sources are acknowledged by explicit references.

I also declare that aspects of the research have been previously published or submitted to journals and conference proceedings. These publications are as follows:

- Gold, E. and Hughes, V. (2014). Issues and opportunities for the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science and Justice*. <http://www.sciencedirect.com/science/article/pii/S1355030614000501>
- Hughes, V., Brereton, A., and Gold, E. (2013). Sample size and the computation of numerical likelihood ratios using articulation rate. *York Papers in Linguistics*, 13, pp. 22-46.
- Gold, E., French, P., and Harrison, P. (2013). Clicking behavior as a speaker discriminant in English. *Journal of the International Phonetic Association,* 43(3), pp. 339-349.
- Aitken, C.G.G. and Gold, E. (2013). Evidence evaluation for discrete data. *Forensic Science International,* 230(1-3), pp. 147-155.
- Aitken, C.G.G. and Gold, E. (2013). Evidence evaluation for multivariate discrete data. In *Proceedings of the 59th ISI World Statistics Congress*, Hong Kong. < http://www.statistics.gov.hk/wsc/IPS021-P1-S.pdf>
- Gold, E., French, P., and Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proceedings of Meetings on Acoustics, (POMA - ICA 2013, Montreal) 19. [DOI: 10.1121/1.4800285]*
- Gold, E. & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law,* 18(2), pp. 293-307.
- Gold, E. & French, P. (2011). An international investigation of forensic speaker comparison practices. In *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, pp. 751-754.*

Signed:

Erica Gold

Date: January 13, 2014

"Not everything that can be counted counts,

and not everything that counts can be counted."

-William Bruce Cameron

# Chapter 1: Introduction

The research presented in this thesis explores the calculation of numerical likelihood ratios using both phonetic and linguistic features. Articulation rate, fundamental frequency, long-term formant distributions, and clicks (velaric ingressive plosive sounds) are analyzed with the purpose of considering intra- and inter-speaker variation, levels of speaker discrimination, the strength of evidence, and the viability of presenting forensic speaker comparison conclusions as numerical likelihood ratios. This chapter outlines the contribution of the thesis in the field of forensic speech science, and provides a short summary of forensic speaker comparison and conclusion frameworks used in forensic speaker comparison cases. The research aims are then described, and overviews of each chapter are provided.

## 1.1 Forensic Speaker Comparison

Forensic speaker comparison (FSC) is noted as being the most common task carried out by forensic phoneticians (Foulkes and French, 2012, p. 558), and the majority of research in the field of forensic speech science (hereafter FSS) is oriented towards this task. FSC is also referred to by other terms such as (forensic) speaker identification, (forensic) speaker recognition, and (forensic) voice comparison (Rose, 2002; Rose and Morrison, 2009). However, the term 'comparison' is preferred in this thesis for two reasons. Firstly, it is not possible to achieve an 'identification' with 100% certainty under a frequentist[1]

[1] The Merriam-Webster dictionary defines frequentist as "[defining] the probability of an event (as heads in flipping a coin) as the limiting value of its frequency in a large number of trials" <http://www.merriam-webster.com/dictionary/frequentist [Accessed 10 January 2014].

conclusion framework, given that there is always, to some extent, within-speaker variability. Secondly, under a Bayesian framework the expert should not take on the role of trier of fact by providing an identification. Rather, his/her responsibility is to express the probability of obtaining the evidence under the hypothesis that the samples came from the same person, versus the probability of obtaining the evidence  under the hypothesis that two different speakers produced the criminal and suspect samples. The term ‘speaker’ is preferred over ‘voice’ in this thesis as not all parameters examined in FSC work are products of just the voice per se. The manifestation of speech parameters can also be a reflection of the social and psychological mind-set of the individual (e.g. French et al., 2010). Therefore, *forensic speaker comparison* is the term preferred over other possible naming conventions.

The analysis in an FSC typically involves the comparison of two (or more) recordings: a criminal sample (also referred to as an ‘unknown’, ‘disputed’, ‘trace’, or ‘questioned’ sample) and a suspect sample (also referred to as a ‘known’ or ‘reference’ sample). The criminal sample is a recording adduced as evidence that contains the speech of an unknown individual. It is possible for the criminal recording to also contain other sounds associated with the crime taking place. In the UK, the suspect sample is usually a recording of a police interview (Nolan, 1983; Rose, 2002) with the suspect.  The objective of the expert forensic phonetician is to provide the trier(s) of fact with an informed opinion regarding the probability of obtaining the evidence (the similarities/differences between the criminal and suspect samples) under the hypothesis that the samples came from the same person, versus the probability of obtaining the evidence (the typicality of the analyzed speech parameters) under the hypothesis that two different speakers produced the criminal and suspect samples. This objective can be reached by experts using a variety of methods (e.g. acoustic analysis, auditory analysis, acoustic and auditory analysis, fully automatic speaker recognition[2] (ASRs), or human-assisted ASR); however, the most common method employed by experts is the combination of auditory phonetic and auditory acoustic analysis of phonetic, linguistic, and non-linguistic speech parameters (e.g. laughter, coughs; French et al., 2010).

### 1.1.1 Expression of Conclusions

Just as there is variability in the methodologies preferred for the comparison of speakers, variability also exists across analyses with regard to the expression of a conclusion at the end of a FSC. Conclusion frameworks can include binary decisions (either the two speakers are the same person or they are different speakers), classical probability scales (probability of identity between the criminal and suspect; Broeders, 1999), the UK Position Statement (a potentially two-part decision based on assessing 'consistency' and 'distinctiveness' of the samples; French and Harrison, 2007), and likelihood ratios (LR, either verbal or numerical, expressing the likelihood of finding the evidence given a same-speaker versus different-speaker hypothesis; Morrison, 2009b; 2009c). There has recently, however, been a strong promotion of the use of the LR framework, as it is advanced as being the only "logically and legally correct framework" (Rose and Morrison, 2009, p. 143).

[2] Other researchers have use ASR to mean automatic speech recognition (e.g. Goel, 2000). However, ASR is used in this thesis to mean automatic speaker recognition.

The use of the numerical LR in research literature has been given increasing attention, starting with Rose (1999). However, the use of the numerical LR in courts in cases involving FSCs is rare (see Rose, 2012; 2013). Many reasons exist for the limited representation of the numerical LR in court, a number of which are addressed in French et al. (2010). However, the key practical limitation cited by French et al. (2010) preventing the implementation of numerical LR conclusions lies in the limited availability of population statistics and the difficulty of collecting them. If numerical LRs were to be calculated with regard to the (very) few parameters for which there are available population statistics, French et al. (2010, p. 149) argue that one “runs the risk of producing an opinion that could lead to a miscarriage of justice.” This is due to the fact that the analysis would fail to consider a large number of other available parameters for which there are no population statistics. In turn, this could impact the conclusion the expert would arrive at in a FSC case.

The motivation for this thesis stems directly from the difficulties and limitations associated with calculating a numerical LR and the increased desire for the field of forensic speech science to align itself with other more developed disciplines of forensic science (e.g. DNA). Previous discussions have tended to focus simply upon the reasons for or against the implementation of a numerical LR. However, far less empirical work has been carried out to examine the practicalities associated with the calculation of numerical LRs. The present study serves as an exercise in calculating numerical LRs for speech data. The data are derived from the speech of a homogeneous group of speakers, while assessing the discriminant ability of these parameters in combination. This exercise is intended to parallel the methodologies and procedures that would be implemented in a real FSC case, should the analyst choose to utilize a numerical LR framework.

## 1.2 Research Aims

The primary aims of this thesis are threefold. The first aim is to provide the field of FSS with a comprehensive summary of current FSC practices used around the world, which has never been previously available. The survey will provide details regarding the types of analysis that are used and their frequency, information on the speech parameters employed, experts' opinions of the discriminant value of those parameters, and the conclusion frameworks they adopt. The survey will also serve as the primary motivation for the selection of the four phonetic and linguistic parameters examined in this thesis.

The second aim is to expand the breadth of FSS literature by examining the actual discriminant value of parameters (individually and in combination) identified by expert forensic phoneticians as being good speaker discriminants. Irrespective of the expectations and actual level of discrimination potential carried by these given parameters, their distributions within the analyzed population will nonetheless provide the field of forensic phonetics with useful information that can further inform FSC casework. The analysis focuses on three intrinsically quantitative phonetic parameters and one ostensibly qualitative linguistic parameter, while aiming to increase the number of speech parameters that can be considered under a numerical LR. Through this analysis, the thesis simultaneously aims to contribute detailed population statistics for four phonetic and linguistic parameters in a large, homogeneous group of speakers. As such, this research will address the arguments set forth in French et al. (2010) about the limited availability of population statistics.

The third, and final, aim of the present body of work is to take the necessary steps to assess the practical limitations and opportunities associated with the implementation of a numerical LR framework in FSCs. This begins by examining potential correlations that exist between and within parameters, and is then followed by appropriately combining the individual pieces of speech evidence. Numerical LRs are then calculated, and strength of evidence and the performance of the combined system are considered. Potential pitfalls and successes are then acknowledged, as doing so is necessary in contributing to the on-going discussion of whether it is practical to adopt a numerical LR framework (in part or full) for FSCs. Most importantly, this work aims to provide a transparent and objective assessment of the viability of implementing a numerical LR framework for FSCs. This assessment will be approached from a structured learning (data-driven) perspective, rather than through giving subjective theoretical opinions regarding the use of numerical LRs in FSC casework.

## 1.3 Thesis Outline

In Chapter 2, an overview is provided of relevant literature relating to the so-called 'paradigm shift' in forensic science, changes in the law, Bayes' Theorem, FSCs, the use of likelihood ratios (LRs) in FSCs, and a discussion of the limitations relating to these topics. The research questions for this thesis that have arisen from previous research reviewed in this chapter are also presented.

Chapter 3 reports on the construction, contents, and results of the first comprehensive international survey of forensic speaker comparison practices. It provides a summary of current practices around the world, commonly-used phonetic, linguistic, and non-linguistic parameters in casework, conclusion frameworks, and expert opinion about which parameters are believed to be highly discriminant. A selection of those parameters found or claimed to be useful speaker discriminants by survey participants is chosen for further examination in subsequent chapters. They include: articulation rate, long-term formant distributions, fundamental frequency, and clicks (velaric ingressive plosives).

The investigation of articulation rate (AR) as a speaker discriminant is presented in Chapter 4. A summary of the AR literature is provided, followed by an analysis of the distribution of AR in the test population. This chapter explores the methodologies used for calculating AR by manipulating the minimum syllable length requirement in a speech interval and comparing inter-pause stretches with memory stretch intervals. The chapter concludes by calculating LRs for AR and determining the levels of discrimination, strength of evidence, and validity of the system.

Chapter 5 analyzes long-term formant distributions (LTFD) as a speaker discriminant. A summary is provided of research that explores LTFD as a parameter for FSC. LTFDs are analyzed individually as well as in combinations relevant to forensic casework, while also providing population statistics. The chapter investigates the effects that the package length (time intervals) of tokens may have on results. Finally, LRs are calculated for individual formants as well as in combination. The levels of discrimination are presented, along with strength of evidence and validity of the system.

Long-term fundamental frequency (F0) as a speaker discriminant is explored in Chapter 6. The chapter offers a summary of the relevant literature on F0 as well as the external factors known to affect F0. Population statistics are offered for F0, and the effects of the package length of tokens are investigated for potential differences in the discriminant results for F0. The chapter concludes with the calculation of LRs, and examines the levels of discrimination, strength of evidence, and validity of F0 as a system.

Chapter 7 analyzes clicks[3] (the final parameter in the thesis) as a speaker discriminant. A summary of the literature on clicks in general is provided at the beginning of the chapter, as well as clicks in conversation analysis. This chapter considers click rate (frequency of velaric ingressive plosives) as a discriminant parameter, and provides population statistics for within- and between-speaker variability. The effects of accommodation are explored in relation to increases in within speaker variation. The chapter concludes by discussing the impeding limitation of not being able to calculate LRs for click rate, due to the lack of appropriate modeling techniques for the data distribution presented by click rate.

The correlations and combinations of those parameters presented in Chapters 4-7 are explored in Chapter 8. The chapter provides a summary of the literature on correlations and combinations of parameters in FSC. Correlations are calculated between all parameters as well as within parameters. These data are used to inform the appropriate methods for the combination of parameters

[3] Used as discourse markers in conversation.

in order to create a complete system (consisting of the four parameters explored in this thesis). Overall likelihood ratios (OLRs) are calculated for the complete system as well as ten alternative systems that consist of different combinations of LTFD, F0, and AR. The performance of the complete system (in terms of strength of evidence and validity) is discussed in comparison to the performance of the alternative systems.

The results presented in Chapters 3-8 are considered collectively and discussed in Chapter 9. A comparative analysis of individual parameters is offered alongside the combination of parameters as a system, examining levels of discrimination between speakers, strength of evidence, and validity. The phonetic-linguistic (human-based) system consisting of AR, LTFD, F0, and clicks is then compared to the performance of ASRs. To conclude, limitations associated with the calculation of numerical LRs are discussed, as well as the implications for using a numerical LR framework in casework.

Finally, Chapter 10 provides a summary of the overall findings of the thesis, revisits the thesis' aims, and identifies opportunities and challenges that face the implementation of a numerical LR should practitioners choose to adopt such a FSC conclusion framework.

# Chapter 2: Literature Review

In this chapter, an overview is presented of the literature surrounding the so-called 'paradigm shift' in forensics, changes in the law, Bayes' Theorem, forensic speaker comparisons (FSCs), the use of likelihood ratios (LRs) in FSCs, and the limitations and shortcomings surrounding these topics that have led to the research questions of this thesis. All subsequent chapters contain a literature review concerning the issue or parameter under focus.

## 2.1 The Paradigm Shift

The term *paradigm shift* was first introduced by Kuhn in 1962. A paradigm in the sciences is defined by Kuhn as a conceptual framework that only members of a particular scientific community share. He goes on to describe a paradigm *shift* as a change in these basic assumptions within the ruling theory of science. An example of one of the most famous paradigm shifts in science is the transition from a Ptolemaic cosmology to a Copernican one, in which the sun is the center of the universe rather than the Earth (Kuhn, 1962). Kuhn (1962) argues that once a paradigm shift is complete, a scientist is unable to reject the new paradigm in favor of the old one. As asserted by Kuhn (1962), paradigms exist in all (sub)domains of science, and forensic science is no exception.

In 2005, Saks and Koehler wrote a review entitled 'The Coming Paradigm Shift in Forensic Identification Science', in which they argued that traditional forensic sciences should "replace antiquated assumptions of uniqueness and perfection with a more defensible empirical and probabilistic foundation" (Saks and Koehler, 2005, p. 895). They begin their review by describing the state of traditional forensic science that follows a frequentist view of evaluation, whereby a decision is made on the probability of a single hypothesis (and without considering prior probabilities). Typically that hypothesis would be that two evidentially-relevant traces, e.g. recorded speech samples, were made by a single object/person (Osterburg, 1969; Stoney, 1991). This form of a hypothesis links evidence to a single object or person to "the exclusion of all others in the world" (Saks and Koehler, 2005, p. 892) Linking evidence to a single object or person is done based on the assumption of uniqueness, whereby the idea is that two evidentially-relevant traces produced by different people or objects will always be different. Therefore when two pieces of evidence are being compared that are not observably different, an expert will conclude that they were made by the same object or person (Saks and Koehler, 2005). The authors are implicitly drawing attention to the single-hypothesis, frequentist-paradigm, in which the typicality of a piece of evidence's characteristics (in a given population) has failed to be taken into account (i.e. they reject the uniqueness assumption).

Saks and Koehler (2005) reveal that in the decade leading up to their review, many people had been falsely convicted of serious crimes, only to be later exonerated by DNA evidence that had not been previously tested at the time of the trial. The authors state that erroneous convictions sometimes occur, and surprisingly in an analysis of 86 cases (ones which resulted in false convictions), it was found that 63% were due *in some part* to erroneous forensic science expert testimony (Saks and Koehler, 2005, p. 892). This was the second biggest contributing factor to false convictions, after misleading eyewitness identifications[4]. The authors explicitly state that the criticism does not apply to DNA evaluation as it is currently practiced; rather, DNA should serve as a model for other forensic science disciplines. The reasoning behind the statement is that DNA typing follows three main principles, (1) the "technology [is] an application of knowledge derived from core scientific principles", (2) "the courts and scientists [can] scrutinize applications of the technology", and (3) it offers "data-based, probabilistic assessments of the meaning of evidentiary 'matches'" (Saks and Koehler, 2005, p. 893).

The authors strongly advise practitioners of other forensic disciplines to emulate the approach taken by DNA typing, whereby the courts are provided with quantifiable evidence, error rates of the technology, and match probabilities being calculated from two competing hypotheses. Without explicitly stating it, Saks and Koehler are essentially arguing for the adoption of the likelihood ratio (LR; § 1.1.1) as the medium for presenting conclusions to the trier(s) of fact (e.g. judge, jury). They are arguing for forensic science to move from a "pre-science to an empirically grounded [one]", that will be transparent and properly scientific. In order for other forensic disciplines to take on such an approach, Saks and Koehler (2005, p. 892) recommend that forensic scientists will need to work closely with experts in other fields to develop efficient methods.

In recent years, following the paper by Saks and Koehler (2005), there have also been calls for improvements in the quality of forensic evidence by a number of legal and government bodies. It has been argued that all areas of

[4] Multiple factors were considered in each false conviction. Therefore, while erroneous forensic evidence was a contributing factor in 63% of the false convictions, erroneous eyewitness testimony was the only other factor contributing to more false convictions (71%; Saks and Kohler, 2005, p. 892).

forensic science need to be more transparent, that forensic examinations should be based on validated methodologies, and that the results should be replicable and expressed in quantitative terms (U.S. National Research Council, 2009; House of Commons' Northern Ireland Affairs Committee, 2009; Law Commission of England & Wales, 2011). These calls for changes to forensic evaluation were made for the same reasons that Saks and Koehler (2005) alluded to with respect to false convictions being made from poorly presented forensic evidence as well as the changes that have occurred in the law.

## 2.2 Changes in the Law

A number of rulings made in the last century have significantly changed the face of expert evidence evaluation and testimony in various countries, especially the United States. Starting in 1923, with the ruling of Frye v. United States, courts moved away from accepting testimony from expert witnesses on the basis of the experts' academic pedigree. Rather, a change was made by a federal appellate court, which rendered expert evidence inadmissible when it was based on methods not used by others in the same forensic discipline. The Frye ruling (Frye v United States (293 F. 1013 D.C. Cir. [1923])) determined that expert testimony was only admissible if the method of analysis used "gained general acceptance in the particular field in which it belongs." (Frye v United States, paragraph 5).

In 1993, the law changed once again as scientific methods continued to improve. In Daubert v. Merrell Dow Pharmaceuticals (509 US 579 [1993]), the United States Supreme Court implemented a new ruling with regard to the admissibility of forensic evidence, whereby the forensic science in question must demonstrate that it can stand on a dependable (i.e. tested) foundation. The ruling challenged those in the field of forensics to show that the forensic method in question had been tested, that its error rate has been established, and this error rate was acceptably low. The Daubert ruling has since been interpreted to mean that forensic sciences should be quantifiable, validated, and reliable. The ruling by the United States Supreme Court was intended to lower the threshold of admissibility for new and cutting-edge methodologies, which would have previously been considered inadmissible under the Frye ruling. At the same time, Daubert was meant to raise the threshold for long-established methods lacking a proper scientific foundation. Daubert subjected forensic sciences to serious methodological scrutiny for the first time.

By 1995, in the case of the United States v. Starzecpyzel, a loophole in the Daubert ruling was brought to light. The case in question included handwriting identification expertise, where a federal district court concluded that handwriting identification had no scientific basis, following the Daubert ruling. This decision was made even though the field of handwriting analysis had dedicated certification programs and professional journals. However, due to the loophole in Daubert the handwriting evidence was not excluded. The reason given was that since the methods used to collect evidence were found to have no scientific basis, Daubert did not apply to handwriting identification as it was not viewed as 'scientific evidence'. The case of Starzecpyzel gave precedent for providers of other forensic testimony to find a way around Daubert by lowering the threshold for admissibility and declaring weakly-founded forensic testimony as non-scientific, thus bypassing the Daubert ruling altogether.

It was not until 1999, in the case of Kumho Tire v. Carmichael, that the United States Supreme Court directly addressed whether or not Daubert applied to ‘non-sciences’. A brief was put together by a number of law enforcement organizations in which they argued that the majority of the expert testimony that they offered did not include scientific theories, methodologies, techniques, or data (Brief Amicus Curiae of Americans for Effective Law Enforcement, 1997). This was stated in relation to the testimony of specific fields of investigation, such as: accident reconstruction, fingerprint, footprint and handprint [identification], handwriting analysis, firearms markings and toolmarks, bullets, and shell casings, and bloodstain pattern identification (Brief Amicus Curiae of Americans for Effective Law Enforcement et al., 1997). Ironically, the practitioners that were initially lobbying for their expertise to be admissible on scientific grounds were now denying that they were a ‘science’. Despite efforts to maintain the ‘non-science’ loophole of Daubert, the United States Supreme Court ruled in Kumho Tire that *all* expert testimony would be required to pass appropriate tests of validity (set forth by Daubert) in order to be admissible in court.

Although the rulings described in this section pertain to law in the United States, these rulings have had a large impact on the legislation in other countries. In the United Kingdom, expert testimony is typically admissible on the basis of the qualifications of the expert testifying rather than the methods employed by that expert. This principle was influenced by the case of R v. Bonython in Australia where the Supreme Court ruled on the admissibility of handwriting evidence. They concluded that forensic evidence testimony is admissible when (i) a layperson is unable to form a sound judgment on the matter “without the assistance of the witness possessing special knowledge or experience in the area”, and when (ii) “the subject matter of the opinion forms part of a body of knowledge or experience which is sufficiently organised or recognised to be accepted as a *reliable* body of knowledge or experience” (R v. Bonython, 1984, paragraph 5). This ruling can be interpreted as being intermediate between the Frye and Daubert rulings, where the Bonython ruling encompasses the Frye ruling and includes the expectation that the testimony is reliable (again, this is usually satisfied with reference to the expert’s academic pedigree and the lack of previous miscarriages of justice in relation to the given expert testimony).

In recent years the House of Commons’ Northern Ireland Affairs Committee (2009) and the Law Commission of England & Wales (2011) have also been influenced by such U.S. rulings, and have urged forensic sciences to make changes to their current practices that would align them more closely with measures set out in Daubert. With respect to the changes in legislature, the developments in presenting DNA typing, the paradigm shift, and the calls made by legal and government bodies, all of these factors (explicitly or inexplicitly) are convergent in their desire for forensic disciplines to adopt a Bayesian framework and to implement likelihood ratios.

## 2.3 Bayes’ Theorem

Bayes’ Theorem was first proposed by Sir Thomas Bayes in the 1740s, then updated and published by Richard Price (Bayes and Price, 1763), and later the same principles were rediscovered and updated further by Pierre Simon

Laplace[5] (Laplace, 1781). Laplace is in reality the man who turned Bayes' Theorem into the modern-day scientific application that is currently used around the world (Bertsch McGrayne, 2012). Bayes' Theorem was created as a way in which to update ones' beliefs. The theorem has three central components: the posterior odds, the prior odds, and the likelihood ratio (LR), as illustrated in Equation 1. The components of Equation 1 are explained in detail in the subsequent sections with respect to forensic science.

(1)

$$\frac{p\left(H_p|E\right)}{p(H_d|E)} = \frac{p(H_p)}{p(H_d)} \times \frac{p(E|H_p)}{p(E|H_d)}$$

*Posterior Odds* *Prior Odds* *Likelihood Ratio*

**Adapted from:** Aitken and Taroni (2004, p. 95)

In Equation (1), *p* represents the probability, where $H_p$ is the prosecution hypothesis (e.g. the criminal and suspect are the same person) and $H_d$ is the defense hypothesis (e.g. the criminal and suspect are different people). The *E* in Equation (1) is representative of the evidence in question. Bayes' Theorem proposes that the posterior odds (the probability of the prosecution hypothesis

[5] The term updated is used here to mean that a prior probability can be adjusted/modified by taking into account any new evidence or observation (e.g. likelihood ratio(s) in forensics) to arrive at a posterior probability (see Equation 1).

being correct given the evidence divided by the probability of the defense hypothesis being correct, given the evidence) is equal to the prior odds (the probability of the prosecution hypothesis being correct divided by the probability of the defense hypothesis being correct; see § 2.3.2 for an example) multiplied by the likelihood ratio (the probability of obtaining the evidence given the prosecution hypothesis divided by the probability of obtaining the evidence given the defense hypothesis; see § 2.3.1 for an example).

### 2.3.1 Likelihood Ratio

The likelihood ratio (LR) is a gradient measure of the value of evidence (Aitken and Taroni, 2004) or what is also referred to as the strength of evidence (Rose, 2002) under a Bayesian framework. An LR is the calculation of the probability of obtaining the results of a given forensic examination on the basis of the prosecution hypothesis divided by the probability of obtaining those same results on the basis of the defense hypothesis. The LR is the only portion of the Bayesian framework in which a forensic expert should provide an opinion. The opinion that the expert provides on the strength of evidence is calculated from two competing probabilities. It should be noted that calculating an LR does not constitute a Bayesian exercise in and of itself (i.e. it only constitutes one part of the Bayesian framework), as that would imply the additional consideration of prior odds (Champod and Meuwly, 2000).

The numerator of the LR is the probability of the prosecutor's hypothesis (the evidence being from the same person/object), while the denominator is typically the probability of the defense hypothesis (the evidence has come from a different person/object). Ideally, the defense hypothesis would be set by the defense (Champod and Meuwly, 2000); however, this is rarely done and the responsibility usually falls to the expert, who typically renders the hypothesis as "the evidence came from someone/something else in the world". The defense hypothesis has also been referred to as a 'random match probability' (Champod and Meuwly, 2000, p. 195). In the LR equation, when the numerator presents a greater value than the denominator, there is support for the prosecution hypothesis, and when the denominator is greater than the numerator, there is support for the defense hypothesis.

The strength or value of evidence in a case is dictated by the magnitude of the resulting LR, or rather the distance of the resulting LR from 1. Therefore, an LR of 100 means that the probability of the evidence (given the competing prosecution and defense hypotheses) is 100 times more likely to have been obtained/to have come from the suspect than someone else in the population. If in the same case the LR was 1/100, then the evidence is 100 times more likely to have been obtained/come from someone in the population other than the suspect (Robertson and Vignaux, 1995). The probabilities of the prosecution or defense hypothesis can take a value between 0 and 1 (inclusive), while the LR can take a value between 0 and ∞ (Aitken and Taroni, 2004). Due to the fact that LRs can be extremely small (approaching 0) or extremely large (tending towards ∞), the LR is often converted into a logarithmic scale with a verbal translation, which makes it easier for the trier(s) of fact (e.g. judge, jury) to interpret (Evett, 1995). If an LR is converted using $\text{Log}_{10}$, a positive value then indicates support for the prosecution hypothesis, while a negative value indicates support for the defense hypothesis.

The combined LRs from multiple pieces of evidence have been referred to as an overall LR (OLR; Alderman, 2004). If multiple pieces of evidence are evaluated in a case, individual LRs can be multiplied together (or added together in the case of $\text{Log}_{10}$ LRs) in order to continue updating an existing probability following Naïve Bayes (Kononenko, 1990; Hand and Yu, 2001). Naïve Bayes refers to when there is an assumption of mutual independence between the pieces of evidence being combined. When cases of correlated evidence (predicted through theory or shown through empirical research) are present, the strength of evidence (resulting LRs) cannot be combined through simple multiplication, and other methods need to be employed. There are at present three general remedies for the problem of combining correlated evidence, (1) the use of an LR algorithm that can handle correlation through statistical weightings (e.g. the Multivariate Kernel Density LR algorithm; Aitken and Taroni, 2004), (2) Bayesian networking that will account for correlations by considering feature distributions and variances and perform statistical weightings (Aitken and Taroni, 2004), or (3) a solution proposed in the field of automatic speaker recognition referred to as logistic-regression fusion, which accounts for correlations in resulting LRs and then applies statistical weightings (Brümmer et al., 2007; Gonzalez-Rodriguez et al., 2007; Ramos Castro, 2007).

In forensics, an LR can be presented either numerically or verbally. An example of a numerical LR statement is, “it is 100 times more probable to obtain the evidence given the prosecution hypothesis than it is to obtain the evidence given the defense hypothesis”. A verbal LR will not include any numbers in its statement; instead, different phrases are used to express the strength of evidence. For example, “it is more probable to that one would obtain the evidence given the prosecution hypothesis than it would be to obtain the evidence given the defense hypothesis”.

### 2.3.2 Prior Odds

In Equation 1, the prior odds represent any existing probability of the hypothesis being true prior to the consideration of new evidence being introduced; they are then updated by the new information, which results in a posterior probability (Aitken and Taroni, 2004). For example, suppose a crime takes place on an island that is inhabited by 101 people. If the perpetrator is known to be one of the 101 inhabiting the island, then the prior odds of the suspect being the criminal is 1/100.[6] These prior odds will then be updated by the trier(s) of fact as new evidence is presented throughout the case. The prior odds are a key factor in the separation between a frequentist way (see § 1.1) of approaching a problem and a Bayesian way of approaching a problem. A Bayesian approach allows the probability of a hypothesis to be updated by any prior probabilities which might affect the posterior probability.

When the prior odds are used in research a numerical value is typically given, and to incorporate those odds into a Bayesian framework it only requires simple multiplication with the likelihood ratio. In practice, the prior odds can be problematic. Robertson and Vignaux (1995, p. 19) state that this is especially true since “very large or very small prior odds can give some very startling effects.” For example, if there were prior odds in a case of ½, multiplied by an LR of 4, the posterior odds would be 2 (i.e. in favor of the prosecution hypothesis). If that same case had prior odds of 1/1000, multiplied by the same LR of 4 (e.g. a

[6] Prior odds = $p(\mathrm{H_p})/p(\mathrm{H_d})$, which in this case is $p(1/101)/p(100/101)$ = 1/100

$H_p$ of 8 divided by a $H_d$ of 2 is equal to 4), the posterior odds would then be 0.004 (i.e. in favor of the defense hypothesis). Significant (or even small) changes of the prior odds can dramatically change the posterior odds. This is demonstrated in the case above, where large prior odds cause the posterior odds to be in favor of the prosecution hypothesis, and much smaller prior odds yield posterior odds in favor of the defense hypothesis (despite the LR remaining constant). Prior odds can also be problematic in practice, given that Bayes Theorem assigns the responsibility of establishing the prior odds to the trier(s) of fact. This means that typically the jurors are held to be responsible for assigning and understanding priors (Robertson and Vignaux, 1995), a task which is in no way simple.

### 2.3.3 Posterior Odds

In Equation 1, the posterior odds are the results of the prior odds after being updated by the LR. Numerically, the posterior odds are the multiplication of the prior odds by the LR (Aitken and Taroni, 2004). Deriving the posterior odds, as with the prior odds, is the responsibility of the trier(s) of fact (Robertson and Vignaux, 1995), and it is up to the trier(s) of fact to determine the posterior odds by considering the prior odds they had initially established, in combination with the evidence provided by expert testimony (the LR(s)). Neither the likelihood ratio nor the prior odds on their own constitute a Bayesian probability; rather, it is the value of the posterior odds that equates to a Bayesian belief of probability.

### 2.3.4 Logical Fallacies

There are three main fallacies that can be committed in the implementation of the Bayesian framework. The first is for the forensic expert to report posterior odds, since the expert does not typically have access to the prior odds, as they are generally set by the trier(s) of fact and not an expert (Rose and Morrison, 2009). Even if the expert were to have access to them, the prior odds will vary in accordance with individuals' personal beliefs about the cases, and the beliefs are subject to natural bias. This fallacy of presenting posterior odds, as committed by a forensic expert, also means that the expert is taking on the role of trier of fact (Robertson and Vignaux, 1995), which in fact infringes on what has been called the 'ultimate issue'. The 'ultimate issue' in law is the decision about the guilt or innocence of a suspect by the trier(s) of fact. If an expert is to present posterior odds, such as an incriminating statement like "the suspect made the shoe mark", the expert then places himself/herself in the role of decision maker, rather than an objective party presenting facts relating to the case (see Joseph Crosfield & Sons v. Techno-Chemical Laboratories Ltd.).

The second fallacy is known as the prosecutor's fallacy (Thompson and Schumann, 1987), also referred to as transposing the conditional (Evett, 1995; Lucy, 2005) or the inversion fallacy (Kaye, 1993). The prosecutor's fallacy occurs when the probability of the evidence given the hypothesis is interchanged with the probability of the hypothesis given the evidence (Lucy, 2005). The inversion of the probabilities gives undue weight to the prosecution hypothesis by assuming that the prior odds of a random match (or two pieces of evidence found to be similar) are equal to the probability of the defense hypothesis. For example, an expert states that there is a 10% chance the suspect "would have the crime blood type if he were innocent. Thus there is a 90% chance that he is guilty" (Aitken and Taroni, 2004, p. 37).

The third fallacy is known as the defender's fallacy (Thompson & Schumann, 1987), which occurs when minimal weight is attributed to the evidence. This is done by considering the background population statistics for a piece of evidence without attention to any associated value (e.g. prior odds). For example, a DNA profile has a probability of 1% in a total population of 10,000, which the suspect comes from. The defense argues that the DNA profile would occur in 100 of these individuals in the population of 10,000, and is therefore of very little value. On the contrary, cutting the total population from 1 in 10,000 to 1 in 100 means that 9,900 people are being excluded, and also it is highly unlikely that all of the 100 individuals are equally likely to be the criminal (Evett and Weir, 1998, p. 32; Lucy, 2005, p. 157).

The three fallacies presented in this section are all flawed in a logical sense as the expert takes on responsibility that is not his/hers (e.g. presenting posterior odds), or the prosecutor/defender only considers a portion of Bayes' Theorem in order to arrive at posterior odds on behalf of the trier(s) of fact. Despite these fallacies having a detrimental effect on the trier(s) of fact's comprehension of the evidence/case with which they have been presented, these fallacies nevertheless need to be monitored in case miscarriages of justice occur. For other errors in the interpretation of Bayes' Theorem, see Aitken and Taroni (2004, pp. 78-95).

## 2.4 Forensic Speaker Comparison

Given the current paradigm shift and the logically and legally correct framework that Bayes' Theorem offers, practitioners in the field of FSC are making the effort to align themselves with other more conceptually advanced areas of forensics (i.e. those using a Bayesian framework, for example, DNA). Acceptance of the forensic paradigm shift in FSC has already been acknowledged and embraced (French et al., 2010). However, the ease with which an LR approach can be adopted is an issue in itself. This is largely due to the challenges that speech data in the forensic context present. This section builds upon the forensic speaker comparison (FSC) introduction in § 1.1. It provides further background information on the complexity of speech data used in FSCs, speech parameters that are commonly analyzed in FSCs, and the way in which FSC conclusions are currently framed in the UK. This section will situate the challenges facing FSC in comparison to other forensic disciplines, while demonstrating the current state of the field as it attempts to align itself with those more advanced (i.e. those using a Bayesian conclusion framework) forensic disciplines.

FSCs are the most commonly performed task by forensic speech scientists (Foulkes and French, 2012). The task of the expert is to provide expert opinion on the speech evidence to the trier(s) of fact. The expert opinion in a FSC is ideally presented in terms of the likelihood of obtaining the evidence (corresponding to the similarities/differences between the criminal and suspect samples) under the hypothesis that the samples came from the same person, versus the probability of obtaining the evidence (the typicality of the analyzed speech parameters) under the hypothesis that two different speakers produced the criminal and suspect samples. The methodologies undertaken by an expert in a FSC are varied, unlike DNA where the same techniques are routinely applied across cases. That is to say that the methodologies involved in FSCs need to be adapted on a case-by-case basis to varying extents. The reasons for this are twofold: (1) speech data is complex in nature (confounding factors are often present), and (2) there is no single speech parameter that is omnipresent and can discriminate all speakers.

### 2.4.1 Complexity of Speech Data

Speech is inherently variable, so much so that phoneticians often make reference to the simple fact that no two speech utterances produced even by the same speaker are ever identical. It is this *intra*-speaker variation that sets forensic speech science apart from some other forensic disciplines. DNA is an example of forensic evidence where the criminal and suspect samples can be identical. For speech, unlike DNA, it will never be the case that the probability of obtaining the evidence given the prosecution hypothesis is ever equal to 1. Variability within the speech of an individual can be caused by numerous factors (e.g. the interlocutor, illness, speaking style, intoxication); however, the maximum extent of variation that can be observed within a speaker is not completely understood through currently available models in linguistics, sociolinguistics, phonetics, or phonology.

The variation that is observed between speakers, or *inter*-speaker variation, is also highly conditioned by both biological and anatomical factors (e.g. vocal tract length, the rate at which vocal cords vibrate), as well as phonological and social factors (e.g. age, sex, class; Chambers, 2005; Eckert, 2000; Wardhaugh, 2006). It is also possible for these variables to interact with one another, whereby their effects are manifested in the speech of individuals differently.

The levels of intra-speaker variation observed in speech recordings are typically high; therefore, it is not surprising that many individual linguistic-phonetic parameters analyzed in FSCs offer only small contributions to advancing the task of speaker discrimination. For this reason, a forensic phonetician will traditionally consider multiple phonetic-linguistic parameters under a combined auditory and acoustic phonetic analysis (French and Stevens, 2013). As a result, the different phonetic-linguistic parameters in a FSC form highly correlated systems and sub-systems owing to the relevant anatomical, phonological, and social factors. The relationships that exist in the data when multiple parameters are under consideration must be appropriately taken into account in the evidence (as is also the case for other forensic sciences). This is so that the conclusion presented in a FSC case is representative of the evidence, and does not over- or under-estimate its strength.

Speech data present a number of challenges to phoneticians looking to analyze phonetic-linguistic parameters in FSCs. This is because the probability distributions associated with phonetic-linguistic parameters are variable. The parameters can be discrete (categorical or qualitative, e.g. impressionistic analysis of voice quality), continuous (e.g. formant frequencies), or a combination of both (e.g. discrete at one level and continuous at another for the same parameter). The continuous parameters are (as a convenient simplification) traditionally assumed to be normally distributed. However, an assumption of normality is not always advised as it can lead to miscarriages of justice (e.g. believing a speaker to be an outlier when he/she is actually very similar to the rest of the population). Algorithms used for calculating LRs (those which assume normality in the data distributions) evaluate similarity and typicality on the basis of data existing at all areas under the normal distribution curve. If a normal distribution curve does not accurately describe the data, the LR algorithms will compute similarity and typicality evaluations from inaccurate descriptions of the data distributions. Additionally, it is possible that the distribution of values of a parameter for an individual speaker is different from the distribution of values of that parameter for a group of speakers.

Finally, in addition to the innate factors that make speech generally complex is the inevitable reality that speech recordings made under forensic conditions are often compromised in terms of quality. Criminal recordings are increasingly recorded via cellular phones, and the recording/transmission technologies involved may affect the quality of the recording. Telephone bandwidth restrictions (Byrne and Foulkes, 2004; Enzinger, 2010b; Künzel, 2001), the distance (of the speaker) from the microphone (Vermeulen, 2009), and cellular phone audio recording codecs (Gold, 2009) have been shown to artificially attenuate portions of the speech signal, which in turn causes unwanted changes to formant frequencies and the fundamental frequency. Furthermore, criminal recordings are also susceptible to low signal-to-noise ratios and high levels of background noise and/or overlapping speech. For this reason, the task of extraction of the necessary parameters is made more difficult in FSC analysis. Despite these problems being prevalent in criminal recordings, they are typically not as severe (or not present at all) in direct and high-quality recordings made in a police interviews (e.g. a recording of a suspect) or other comparable situations.

Given that the suspect recording in a FSC case typically comes from a police interview, there is often a mismatch in the conditions under which the criminal and suspect samples are elicited. The criminal recording is frequently made in situations that involve high emotional states, physical activity, or the influence of drugs or alcohol. They also tend to be short in duration and limited in content. This presents the forensic phonetician with additional complications.

### 2.4.2 The Phonetic Shibboleth

The search for the linguistic or phonetic shibboleth[7] for discriminating speakers has proved fruitless since research began in the field of FSS. This should not come as a surprise, given the inherent complexity of speech data, as outlined in § 2.4.1. Research has shown that the vocal tract is highly plastic, that no phonetic/linguistic parameter is omnipresent, that a phonetic/linguistic parameter that makes one speaker different does not necessarily make another speaker different, and that parameters that make a speaker differ can vary over time. It is likely, furthermore, that it is the combination of parameters that makes a speaker unique (Nolan, 1983; Rhodes, 2013; Rose, 2013a).

There is a large and growing body of literature devoted to identifying phonetic and linguistic parameters that have ideal characteristics for FSC. These criteria have been outlined by Nolan (1983, p. 11):

[7] Shibboleth is used here to refer to a single identifiable parameter that can discriminate between all speakers.

1. High between-speaker variability: The parameter should show a high degree of variation between speakers. If a single parameter cannot show this then a set of parameters can be sought

2. Low within-speaker variability: The parameter should show consistency throughout the speech of an individual, and be insensitive to external factors (e.g. health, emotion, or interlocutor)

3. Resistance to attempted disguise or mimicry: The parameter must withstand attempts on the part of the speaker to disguise his voice

4. Availability: Any parameter should provide an ample amount of data in both the criminal and suspect samples

5. Robustness in transmission: The usefulness of a parameter will be limited if its information is lost or reduced due to recording or transmission technologies

6. Measurability: The extraction of the parameter must not be prohibitively difficult

Criteria 3-6 are specifically concerned with practical issues that arise in casework. Criterion 4 relates to the often limited amount of material an expert is given to work with, while criteria 5 and 6 are associated with the recording and transmission technologies typically used in forensic recordings. The first two criteria suggested by Nolan (1983) identify the true difficulty of the FSC task. That is, the expert has to identify and examine phonetic and linguistic parameters that have high inter-speaker variation, but also low intra-speaker variation. It is often the case that a phonetic or linguistic parameter meets a single criterion, but a phonetic/linguistic shibboleth is yet to be found that meets both criteria without exception.

#### 2.4.2.1 Research Question 1

Given the difficulties and limitations in selecting highly discriminant phonetic and linguistic parameters for analysis in FSCs, the most obvious question is:

(1) What phonetic and linguistic parameters do practicing forensic phoneticians (around the world) typically analyze in a FSC case and which parameters do they view as being highly discriminant?

Before questioning the proper (or logically and legally correct) framework in which to make conclusions about FSC evidence, it is important to consider the methodologies, practices, and parameter selection that is involved in the actual FSC analysis itself. Only after establishing the general expectations of the FSS community should one begin to broach the problem of FSC conclusion frameworks. For without any analyzed forensic speech evidence, there can be no valid conclusion.

### 2.4.3 Current Conclusion Framework in the UK

The current practice for presenting FSC conclusions in a UK court is not in the form of an LR as described in § 2.2, but rather is that described in the UK Position Statement that was introduced in 2007. The UK Position Statement was motivated by concerns about "the framework in which conclusions are typically expressed in forensic speaker comparison cases" (French and Harrison, 2007, p. 137). The UK Position Statement stemmed from ruling of the Appeal Court of England and Wales in R v. Doheny and Adams (1996), which showed that the interpretation of the DNA evidence at the initial trial had been flawed by the prosecutor's fallacy (French and Harrison, 2007). The introduction of the UK Position Statement signified a shift in the role of the forensic phonetician when presenting speech evidence. The foreword to the UK Position Statement suggests that experts in the past were often trying to identify speakers (French and Harrison, 2007, p. 138). However, under their new approach an expert would not be making identifications per se. Instead, the expert will take on a different role (not one of speaker identification), to provide "an assessment of whether the voice in the questioned recordings fits the description of the suspect" (French and Harrison, 2007, p. 138). The UK Position Statement was also proposed with the intention of aligning the field of FSC with "more modern thinking" forensic sciences (French and Harrison, 2007, p. 137).

The framework laid out in the UK Position Statement diverges from previous FSC conclusions by offering a framework which involves a bipartite assessment. The conclusion framework set out in French and Harrison (2007) potentially involves a two-part decision. The first part concerns the assessment of whether the samples are consistent with having been spoken by the same person. The second part, which only comes into play if there is a positive decision concerning consistency, involves an evaluation of how unusual or distinctive the combination of features that are common to the samples may be. An illustrated version of the UK Position Statement is provided in Figure 2.1.

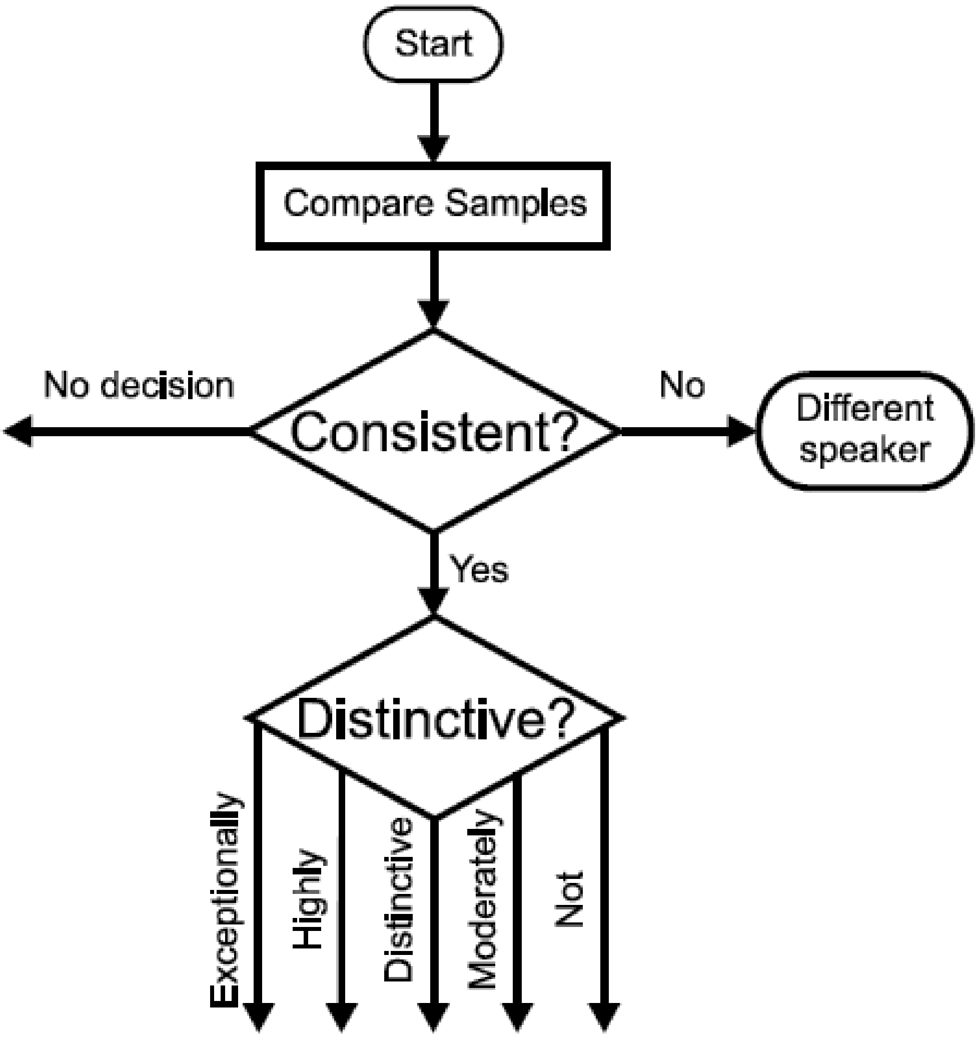


**Figure 2.1:** Illustration of the UK Position Statement (Rose and Morrison, 2009, p.141)

The UK Position Statement is illustrated in Figure 2.1, where the decisions of consistency and distinctiveness are serially ordered. A consistency decision has three possible options: consistent, not-consistent, and no-decision. If a conclusion about consistency cannot be made, then the expert concludes with a single evaluation (i.e. not consistent, or no decision). In the event that the expert finds the two speech samples to be consistent, s/he will then assess the distinctiveness. The degree of distinctiveness is made on a five-point impressionistic scale, ranging from “not distinctive” to “exceptionally distinctive”. The assessment of distinctiveness in many cases must draw upon the experience of the expert so that he/she can provide a statement of the typicality of the criminal speech sample.

The UK Position Statement framework can be seen as a transitional point, or a stepping stone, between a frequentist probability  and an LR, where it is not providing a single probability of the hypothesis (e.g. the speaker in the criminal sample is likely to be person X), but not quite meeting the logical framework of the LR. At first glance the judgments of consistency and distinctiveness appear to mirror the numerator and denominator of an LR, as the consistency and distinctiveness account for both the similarity and the typicality of the speech recordings. However, the inner workings of the Position Statement do not hold true to the logical framework of an LR. There are two main reasons for this mismatch: (1) assessments are made on different scales, and (2) there is no logical procedure for combining (and weighing) constituent speech parameter evidence from a single case.

Rose and Morrison describe the assessment of consistency in the UK Position Statement as being on a three-point scale (Rose and Morrison, 2009, p. 142). Although Rose and Morrison acknowledge that the decision about consistency is categorical, one could argue that the assessment of consistency is more accurately described as simply a ternary decision (rather than on a three-point scale). This is due to the fact that the judgment is wholly categorical, and the ternary decision cannot be intuitively placed on a scale. A scale would imply some degree of hierarchy, and it is difficult to argue that, for example, no-decision should be ranked before inconsistent (or vice versa). Therefore, the assessment of consistency is discrete in nature and does not offer a gradient assessment of the similarity (through quantification of the speech evidence), as the numerical LR would ultimately provide. The assessment of distinctiveness is on a scale of one to five; however, this does not follow the same logic as the assessments of consistency. Thus, it is difficult to establish a working relationship between the two assessments; instead they exist more as two separate entities, where practitioners are trying to make a judgment on the same piece of evidence.

The use of a five-point scale in the UK Position Statement makes the framework prone to a cliff-edge effect (Aitken and Taroni, 2004). By imposing defined boundaries an expert is faced with a hard decision. So, for example, if a criminal sample has an F0 mean of 115 Hz, while the population mean is 90 Hz, should the analysis of a speech sample lend itself to a distinctiveness assessment of 3 (distinctive) or 4 (highly distinctive)? Should the boundary between distinctive and highly distinctive be two standard deviations, or perhaps three? Forcibly imposing categorical boundaries could potentially over- or under-estimate the strength of evidence. Although the UK Position Statement is susceptible to the cliff-edge effect, the same can actually be said for the verbal LR scale provided by Evett (1998). Although the verbal scale suggested by Evett (1998) is associated with $\text{Log}_{10}$ LRs, the cliff-edge effect can still occur for those $\text{Log}_{10}$ LRs that lie close to the categorical boundaries.

The second inconsistency between the UK Position Statement and the LR is the lack of a protocol for combining the strength of evidence of individual phonetic-linguistic parameters. Under a Bayesian framework an expert is expected to combine individual LRs for parameters that are mutually independent (Kononenko, 1990) by multiplying their LRs. If an expert is to naïvely combine correlated parameters without using appropriate statistical weightings, s/he then runs the risk of over-or under-estimating the strength of evidence (as s/he are essentially considering the same evidence multiple times). Under the UK Position Statement, experts make assessments of consistency and distinctiveness by informally considering all of the constituent pieces of analysis together. As such, they are unlikely to adequately (or transparently) consider the degree of correlation between the evidence. Therefore, conclusions made under the UK Position Statement framework could over- or under-estimate the strength of evidence.

Despite the disparity between the UK Position Statement and the LR, the UK Position Statement possesses two highly attractive attributes. Firstly, it allows the expert to avoid the undesirable and lengthy task of collecting (quantitative, data-based) population statistics for all possible relevant populations that could ever be required for a FSC case. Secondly, the framework allows the expert to avoid the difficulty of calculating numerical LRs for all phonetic-linguistic parameter distributions that do not fit into already existing LR algorithms. These two attributes are technically part and parcel of the same thing, as they together evaluate the denominator of the LR; however, the modeling of phonetic-linguistic parameters is also pertinent to the numerator. It is safe to argue that no LR algorithm could ever account for or encompass the full complexity of speech data; therefore, perhaps the UK Position Statement is right to circumvent fully quantitative population statistics and complicated models for calculating LRs for all speech parameters. Through experience and education, an expert is able to account for instances of accommodation, channel mismatch, intoxication, emotional effects, and social factors. These factors tend to manifest themselves differently in the speech of each individual speaker and at different times. To create an algorithm that accounts for every individual, in every instance, would be near impossible.

## 2.5 Likelihood Ratios in Forensic Speech Science

It is now generally accepted in forensics that the logically and legally correct framework for expressing the results of forensic examinations is one in which the output is a likelihood ratio (Saks & Koehler, 2005). The domain of forensic speech science is no exception to this, and efforts have been made in the last decade and a half (starting with Rose, 1999) to incorporate into the LR into forensic phonetic- (and linguistic)-based research and casework. In forensic phonetics, the LR essentially becomes a test of the similarity and typicality of phonetic-linguistic parameters that are extracted from recordings. The numerator of the LR contains the probability of obtaining the evidence given the hypothesis that the speech came from the *same speaker*, while the denominator is the probability of obtaining the evidence given the hypothesis that the speech came from a *different speaker* (Rose, 2002). The same-speaker hypothesis is determined by comparing speech parameters from the criminal and suspect samples to establish the degree of similarity. The different speaker hypothesis is determined by comparing speech parameters from the criminal speech sample to those drawn from a relevant background population so as to establish the degree of typicality. The probability obtained from the numerator is then divided by the probability obtained from the denominator, and the result is the LR for the given speech evidence.

The presence of LRs in FSC is largely confined to the research literature with only a single case example of the LR being used in FSC casework to date (Rose, 2012; 2013). The research carried out has predominantly had two main foci: (1) assessing speaker discrimination using a numerical LR and (2) overall improvements in LR methodologies. The next two sections (§ 2.5.1 and § 2.5.2) provide an overview of methodological research that has been carried out, as well as a review of the application of LRs in practice.

### 2.5.1 Likelihood Ratios in the Literature

This section focuses on the LR literature that investigates the use of numerical LRs as a framework for carrying out the assessments of speaker discrimination ability using phonetic and linguistic parameters, and the LR literature that seeks to improve current methodologies.

#### 2.5.1.1 Likelihood Ratios for Speaker Discrimination

The application of the LR framework to FSCs has focused almost exclusively on vowels. Vowels can be easy to extract quantitative measurements from, and so readily lend themselves to the calculation of numerical LRs. In order to improve discrimination rates between speakers, researchers have measured vowels with multiple methodologies: using mid-point formant values (Alderman, 2004; Rose, 2007a; Rose, 2010a; Rose and Winter, 2010; Zhang et al., 2008), formant trajectories of monophthongs and diphthongs (Atkinson, 2009; Enzinger, 2010a; Kinoshita and Osanai, 2006; Morrison, 2009a; Rose et al., 2006), and long-term formant distributions over vowel mixtures (Becker et al., 2008; French et al., 2012; Moos, 2010).

Comparatively speaking, non-vowel research for speaker discrimination purposes has not been given the same amount of attention as vowel-based LR research. Those studies that have been carried out on features other than vowels have all focused on quantitative, multivariate data that is typically normally distributed. The non-vowel parameters that have been investigated include fundamental frequency (F0), voice onset time (VOT), nasals, laterals, and fricatives (Kavanagh, 2010; 2011; 2013; Kinoshita, 2002; 2005; Kinoshita et al., 2009, Coe, 2012). Traditional FSC does not *just* involve the analysis of vowels and the non-vocalic features listed above. For this reason, further empirical work is required to evaluate the discriminatory value of additional speech parameters using a numerical LR.

#### 2.5.1.1.1 Research Question 2

The current body of literature evaluating the discriminant ability of speech parameters is plentiful. However, there are a number of speech parameters that have not had their discriminant ability tested. To date, parameters have been selected for analysis based principally on their ease of measurement. Instead, it is proposed here that parameters should be selected on the basis of their discriminatory merit as proposed on the basis of the experience of forensic phoneticians. These considerations lead us to ask our second research question:

(2) If experts are to provide their opinions on the most helpful speaker discriminants, are these 'selected' parameters going to be good speaker discriminants?

   a. Furthermore, do experts' expectations surrounding the discriminant value of certain speech parameters match the results of these parameters' empirically-tested performance?

A simple hypothesis to test when empirically evaluating parameters that are identified as being commonly used in FSCs and which experts propose to be useful discriminants is that such parameters will perform better than speech parameters selected for analysis arbitrarily (simply because they are easily measurable and plentiful).

#### 2.5.1.2 Improving Likelihood Ratio Methodologies

In addition to the LR literature that has assessed the discriminant ability of phonetic parameters, there is a dedicated body of literature on methodological advances in the calculation of LRs in other domains. In particular, there have been methodological advances across a range of areas, including the development of modeling techniques of data for calculating LRs (Kinoshita, 2001; Morrison, 2011; Zhang et al., 2008), exploring the issues surrounding correlated parameters (Gold and Hughes, 2012; Morrison et al., 2010; Rose, 2006c; 2010b; Rose et al., 2004), identifying the relevant population for the LR (Hughes, *in progress*; Hughes and Foulkes, 2012; Morrison et al., 2012a; 2012b), exploring the amount of data that is preferred for the reference population (Hughes, *in progress*; Hughes and Foulkes, 2012; Ishihara and Kinoshita, 2008; Kinoshita and Ishihara, 2012), combining parameters

(Morrison et al., 2010; Morrison, 2013; Rose 2010a; 2010b, 2013a; Rose et al., 2004; Zhang et al., 2008), system calibration (Morrison, 2012; Morrison et al., 2010; Morrison and Kinoshita, 2008), and measures of system validity and reliability (Morrison et al., 2010; Morrison and Kinoshita, 2008).

Despite these methodological developments, there are numerous research questions relating to the calculation of LRs in FSCs that would greatly benefit from further empirical study and assistance from forensic statisticians. The majority of the previous research has perhaps neglected to acknowledge the complexity of speech data and has opted for often convenient but erroneous simplifications of basic linguistic principles in order to calculate LRs.

#### 2.5.1.2.1 Research Question 3

As previously outlined in § 2.4.2, Nolan (1983) recommended that a *set of parameters* should be sought to show high between-speaker variability where a single parameter alone is not sufficient. Given the large body of literature on single speech parameters as discriminants and their limited discriminant power, it is suggested that further work needs to heed Nolan's suggestion.

(3) How well do speech parameters work in combination to discriminate between speakers?

   a. What steps need to be taken in order to appropriately combine speech parameters?
   b. Is the combination of multiple speech parameters always better than individual parameters at discriminating between speakers?

One would hypothesize that adding ever more parameters would further advance the task of FSC, since theory and research tells us that speakers are different from one another in a variety of ways. By combining multiple speech parameters, it is proposed that a combined system will achieve better discrimination performance than those achieved by single parameters.

### 2.5.2 Likelihood Ratios in Practice

The only publications that report the use of LRs for multiple speech parameters in FSC casework are those of Rose (2012; 2013b) in connection with a fraud case in Australia. The case of R v. Hufnagl (2008) revolved around a large-scale telephone fraud of AUS$150 million, where a criminal sent a fax to JP Morgan Chase bank, asking to transfer $150 million from the Australian Commonwealth Superannuation Scheme to accounts in Switzerland, Greece, and Hong Kong. Before the close of business, the criminal called the bank asking for confirmation of the details in the fax he had sent. When the Australian Commonwealth Superannuation Scheme realized their account was short by $150 million, an investigation followed. A suspect was identified, and Rose was asked to compare the recording of the fraudulent telephone call with recorded telephone calls known to have been made by the suspect. The analysis and report were produced five years prior to Rose's publications about it (2012; 2013b), so he presents the original analysis that was carried out as well as a retrospective critique of his analysis.

In the original analysis, he identified many tokens of the word *yes* in both the criminal and suspect recordings, as well as the utterance *not too bad* in the criminal recording and multiple occurrences of the same phrase in the suspect recordings. Therefore, the majority of the analysis and the resulting LR were based on phonetic/linguistic parameters measured from these words. In order to establish the typicality of the criminal's speech, Rose defined the relevant population to be "adult male speaker[s] of General Australian English" (Rose, 2013b, p. 284). He then collected relevant speech samples from 35 adult males, who served as the background population. The analysis of similarity was comprised of formant measurements from /je/ in the word *yes* at three designated time-points, the fundamental frequency (F0) in *not too bad* taken from four designated time-points, categorical classification of high and low tones in *not too bad*, formant measurements of the vowels in *not too bad*, and the frequency cut-off in /s/ from the word *yes* (Rose, 2013b). After (intentionally) naïvely combining the individual LRs from the parameters, an overall LR (OLR) of around 11 million was calculated. Rose (2013b) explicitly states that 11 million was an over-estimation of the strength of evidence, since some degree of correlation had to exist between the parameters. For this reason, parameters that were assumed to have some degree of correlation with one another (e.g. formant measurements for certain vowels) were thrown out, and a more conservative LR of 300,000 was reached.

Five years after the conclusion of the case, Rose provided a critique of the analysis and presentation of the evidence under an LR framework. He notes a number of developments made in the field since the R-v-Hufnagl case that could have made a significant difference in his analysis. These include vowel (and consonant) parameterization (e.g. formant dynamics; McDougall, 2004), quantification of accuracy and precision (validity and reliability; e.g. Cllr and EER for validity measures), and - most importantly - techniques to handle between-parameter correlations for calculating OLRs (e.g. fusion). If any of these developments were to have been implemented in R v. Hufnagl (2008), it can confidently be said that the strength of the numerical LR would not be identical to that presented in Rose (2013b; also shown through his reanalysis of the case material); most likely, the strength of the LR would weaken as correlated parameters were accounted for during the combination of speech evidence (acknowledged in Rose, 2013b).

The final portion of Rose (2013b) commented upon the court's reaction to the presentation of evidence in the form of a numerical LR, which is something rarely discussed in forensic phonetics. The expert testimony did not include a complete tutorial on the LR approach; rather, it offered a more abstract presentation of the strength of evidence (the LR). Rose (2013b) condensed his analysis into two main points for the jury, which he emphasized on multiple occasions: (1) the LR is for estimating the strength of the evidence and not the probability that the suspect is the criminal, and (2) the jury should not give much weight to the specific value allocated to the LR, just that it was very big. Whether Rose's testimony made an impression on the triers of fact in R-v-Hufnagl is unknown. However, the jury did return a guilty verdict (Rose, 2013b). Rose also notes that it was perhaps vital to his testimony that the judge was encouraging towards his approach and that this helped him (Rose) to articulate to the court the strength of the speech evidence. It can be assumed that not all judges would act in the same manner, and presenting the same testimony in front of a different judge might have been more challenging without such support.

Overall, it is encouraging to see an example of a real case in which a numerical LR framework was used. The introduction of Rose's paper provides a nice backdrop to the case and the type of speech material Rose chose to analyze. The critique at the end of the paper is a positive contribution, as it shows how the field has evolved in the past five years since the case analysis was completed. The paper also shines light on the reception of the LR in a court, which again often goes without attention in the literature. However, the paper perhaps brings up more questions (both theoretical and practical) about the implementation of the LR framework (as used by Rose) than it answers. For instance, how does an expert begin to select parameters for analysis under an LR framework? How can an expert argue why s/he has selected certain parameters for analysis over other parameters? How is an expert to incorporate qualitative/categorical parameters? And how many parameters need to be analyzed to consider the evaluation to be complete?

Despite raising a new set of questions, Rose (2013b) makes three pertinent statements with respect to LRs. These statements are particularly relevant to the remainder of this thesis. The first is that "real-world cases are never the same" and "there is no one-size-fits-all" with regards to methodology (Rose, 2013b, p. 318). This means that the LR calculation is not the same in every FSC case, or for every phonetic/linguistic parameter selected for analysis. Therefore, the analysis that leads to an LR will always have to be adapted on a case-to-case basis. The second statement asserted by Rose is that FSC might lend itself more readily to a verbal LR over a numerical LR[8]. The reason for this

[8] A verbal LR is simply a verbal, rather than numerical, statement of the probability of obtaining the evidence given the prosecution hypothesis over the probability of obtaining the evidence

is that precise figures may be misleading in that numerical LRs may be difficult for the trier(s) of fact to interpret[9] (Rose, 2013b, p. 305). The final statement comes from Judge Hodgson (2002) but is reiterated by Rose (2013b): "since not all types of evidence in a trial can be sensibly assigned a LR there is no way of mathematically combining à la Bayes the LR-based evidence with the non-numerically based evidence" (Rose, 2013b, p. 316-317). This leaves one to ponder whether there is really an explicit need for speech evidence to be represented in numerical LR form. For example, would a phonetician ever be able to quantify the exact tongue shape of a speakers' /ɹ/? In this instance, a qualitative description of /ɹ/ will typically be more useful than a quantitative one that is not completely transparent in its description. Should these types of evidence always be unsuitable for expression in a numerical LR, will it be the case that other phonetic-linguistic parameters can be made to fit the mold in the form of LR algorithms that dictate specific quantitative forms? It is also important to consider that if a numerical LR is used, only a partial assessment of the speech evidence is feasible, given that numerical LRs cannot currently be calculated for all speech parameters (because of the lack of appropriate algorithms and/or the qualitative nature of certain parameters), and the lack of population statistics in general.

#### 2.5.2.1 Research Question 4

The literature review provided in the previous sections revealed a number of limitations and difficulties that can occur when applying the

given the defense hypothesis. For example, the verbal statement could be presented as 'it is extremely more probable to obtain the given evidence under hypothesis x than y.'

[9] For example, is there really much of a difference between an LR of 1.1 x $10^{14}$ and an LR of 1.11 x $10^{14}$?

numerical LR framework to FSCs, which are largely due to the complexity of speech data. If the field is to continue in its efforts to align itself with more advanced forensic disciplines (in terms of conclusion frameworks) that have already adopted the LR framework (e.g. DNA), various aspects of the actual calculation of an LR in a FSC should be reviewed and improved (e.g. modeling techniques, population statistics, combining parameters for OLRs).

(4) For this reason, it is essential to ask: What are the practical limitations/implications that need to be considered when using the numerical LR framework in FSCs?

   a. What recommendations, if any, can be provided following attempts to implement the numerical LR framework?

   b. What can a human-based (acoustic-phonetic) system tell the field in respect of the ease with which a numerical LR can be computed for FSCs?

The practical limitations and implications associated with the implementation of a numerical LR will be discussed throughout this thesis. It is only through empirical testing that these questions can be addressed.

## 2.6 Summary of Research Questions

This chapter has presented a series of research questions that have been motivated by the prior literature and existing legal rulings with regard to forensic evidence, while further developing the research aims of the thesis. This section reiterates the research questions identified in this chapter, which will be explored in the remainder of the thesis.

(1) What phonetic and linguistic parameters do practicing forensic phoneticians (around the world) typically analyze in a FSC case and which parameters do they recommend as being highly discriminant?

(2) If experts are to provide their opinions on the most helpful speaker discriminants, will these 'selected' parameters be good speaker discriminants?

   a. Furthermore, do experts' expectations surrounding the discriminant value of certain speech parameters match these parameters' empirically-tested performance?

(3) How well do speech parameters work in combination to discriminate between speakers?

   a. What steps need to be taken in order to appropriately combine speech parameters?

   b. Is the combination of multiple speech parameters always better at discriminating between speakers? Are more parameters better?

(4) What are the practical limitations/implications for using the numerical LR framework in FSCs?

   a. What recommendations, if any, can be provided following attempts to implement a numerical LR framework?

   b. What can a human-based (acoustic-phonetic) system tell the field in regards of the ease with which a numerical LR can be computed for FSCs?

# Chapter 3: International Survey of Forensic Speaker Comparison Practices

## 3.1 Introduction

This chapter presents the results of the first comprehensive international survey on forensic speaker comparison (FSC) practices.

The motivations for the survey were twofold:

(i) For the first time, to make available to the wider forensic, legal, and speech science communities basic information concerning the working practices of FSC experts around the world.

(ii) To draw upon the very considerable collective experience of FSC experts worldwide in order to identify current working methods and features of speech that are considered to have the greatest potential for discriminating between individuals.

It will become apparent from the results presented below that there is a great deal of variation in the methods of analysis, features selected for examination, weighting attached to certain features relative to others, and frameworks used for expressing the conclusions that arise from the comparisons. Some of the differences found are, undoubtedly, a function of the rules, regulations and laws of the institutions and jurisdictions in which the survey participants are working. Others, however, would appear to be simply a matter of local tradition or individual intellectual preference. The results are therefore discussed in the context of the constraints on the admissibility of expert evidence in different countries and are related to contemporary debates within forensic speech science.

### 3.1.1 Background

While research has been carried out on many facets of FSS and FSCs, there has not been any research that has comprehensively surveyed the FSC practices employed by experts around the world. The extent to which the FSS community had been aware of commonly used FSC practices has been limited to the results of an exercise Cambier-Langeveld (2007) conducted using a fictional FSC case. The objective of the exercise was not to survey practitioners, but rather to observe and assess basic methods that participants chose to employ in conducting the fictional FSC case. Cambier-Langeveld's paper considers reports from 10 of 12 participants based in 10 different countries. Her article reports on some of the basic methods involved in a FSC case, which were confined to: the length of recordings needed for speech samples, formant measurements, fundamental frequency, and the formulations of conclusions. The results of the exercise revealed inconsistencies in methods amongst the 10 participants. However, it usefully relayed fundamental methodological information with regard to FSCs that was previously unavailable.

The exercise conducted by Cambier-Langeveld (2007) was an attempt to provide the field of FSS with a body of information relating to FSC methods, but not to provide a wide-ranging picture of current practices. However, the study did create interest and a platform on which to conduct further research into the methodologies employed by the field for FSCs worldwide.

Hollien and Majewski (2009) discuss the prevalence of inconsistencies in FSS practices with particular attention to FSCs, like those reported by Cambier-Langeveld (2007). The authors suggest that the field of FSS lacks any real consensus in terms of procedures and methods for FSC cases. They argue that it is difficult to consider what level of scientific probability is robust enough to determine the identity of a speaker, and that without any standards or common practices it is difficult to make comparisons across different approaches. They offer a protocol for a frequentist conclusion framework that they implement (and that other experts could adopt should they wish to), which includes confidence levels of their judgments. Despite their efforts to offer their own standard and protocol for FSCs, the authors fail to acknowledge alternative methodologies that are currently being implemented by experts in FSC cases across the globe. I would argue that an understanding of the current state of the field is a prerequisite for establishing any form of standards or protocols.

For a field that came to fruition in the late 1980’s/ early 1990’s (the time at which acoustic and auditory phonetic analysis began regularly being used in the UK courts at least, (French, p.c.)) little has been done to unify and standardize the field over this time. While Cambier-Langeveld (2007) and Hollien and Majewski (2009) argue that there is a lack of consensus in the field of FSC and that standards are almost non-existent, I would suggest that the only way to remedy such a fault is first to assess the current methodologies and practices being used in FSCs by surveying expert forensic phoneticians around the world.

## 3.2 The Survey

The survey was administered online using SurveyGizmo 3.0. It consisted of 78 questions related to all aspects of forensic speaker comparison casework. All participants were kept anonymous and also given the option of answering some or all questions. Although every question was answered by at least 30 participants, the variability in respondent numbers nevertheless dictates that the majority of the results must be presented as percentages.

### 3.2.1 SurveyGizmo

"SurveyGizmo is an online survey software tool for designing online surveys, collecting data and performing analysis. [The] tool supports a variety of online data collection methods including online surveys, online quizzes, questionnaires, web forms, and landing pages" (SurveyGizmo, 2010). It was selected as the medium for the survey over other similar websites for a number of reasons: the server is secure, the package offers the ability to save and continue (when taking the survey), an inexpensive Student Account with enhanced privileges, and an excellent user interface for creating the survey. All responses collected from the survey are saved on the SurveyGizmo server, and answers can only be accessed by a username and password.

### 3.2.2 Methodology: Data Compilation

To complete the survey, participants were provided with a survey link in an email invitation. They were then redirected to the SurveyGizmo website where they gave their consent to participate in the survey and agreed to their data being used in future research. After giving consent, participants were provided with instructions (see Appendix A) as well as an outline of the survey structure. They were allowed to stop the survey at any time and save it, so that it could be completed at a later time. Many of the respondents took advantage of this feature as the total time (including interruptions) it took most participants to complete the survey ranged from 26 minutes to 64 hours.

Once all participants had submitted their answers to the survey questions, the results were tabulated using Microsoft Excel.

## 3.3 Participants

Potential participants were contacted through their professional and research organizations. Emails were sent to the European Network of Forensic Science Institutes, the National Institute for Standards and Technology (NIST) for those who participate in the NIST Speaker Recognition Evaluations, and the International Association for Forensic Phonetics and Acoustics. Some individuals working at government laboratories/agencies were contacted through their employers. In total, 36 practicing forensic speech scientists agreed to participate, and data were collected from July 2010 through March 2011.

### 3.3.1 Countries

Respondents (23 male; 13 female) were from the following 13 countries: Australia, Austria, Brazil, China, Germany, Italy, the Netherlands, South Korea, Spain, Sweden, Turkey, UK, and USA. Although the majority of the participants were from Europe, a total of five continents were represented in the results.

### 3.3.2 Place of Work

Respondents identified their place of work[10] or affiliation. 18 participants represented universities or research institutes, followed by 13 who were employed in government laboratories/agencies. 9 of the experts are affiliated with private laboratories, and 7 work as individuals.

### 3.3.3 Experience

The total number of cases from respondents' collective estimates was 18,221, ranging from 4 to 6,000, with a mean of 506. The respondents had a range of 2 to 50 years of experience in FSC analysis, with a mean of 15.

## 3.4 Methods of Analysis

Participants' responses showed that there is at present no consensus of opinion in the scientific community as to how FSC analysis should be carried out. Rather, a wide range of methods is employed. Methods may be grouped under the following headings:

***Auditory Phonetic Analysis Only (AuPA):***

The expert listens analytically to the speech samples and attends to aspects of speech at the segmental and suprasegmental levels.

***Acoustic Phonetic Analysis Only (AcPA):***

The expert analyses and quantifies physical parameters of the speech signal using computer software. As with AuPA, this is labor-intensive, involving a high degree of human input and judgment.

[10] Some respondents are associated with multiple places of work.

***Auditory Phonetic-cum-Acoustic Phonetic Analysis (AuPA+AcPA):***

This combines the preceding two methods.

***Analysis by Automatic Speaker Recognition System (ASR):***

This requires the use of specialist software designed to estimate the degree of similarity between speech samples based on statistical models of features extracted automatically from the acoustic signal. Such systems typically require minimal input from the analyst.

***Analysis by Automatic Speaker Recognition System with Human Analysis (HASR):***

This involves the use of an automatic system in conjunction with analysis of the auditory and/or acoustic phonetic kind. The survey did not investigate the precise nature or extent of the auditory or acoustic examinations experts used to supplement the ASR component. The human-based supplementary analysis may range from cursory holistic listening to detailed auditory and/or acoustic examinations.

More detailed descriptions of these methods, either individually or relative to one another, may be found, *inter alia*, in Baldwin and French (1990), Drygajlo (2007), French (1994), French and Stevens (2013), Greenberg et al. (2010), Jessen (2007a; 2008), and Künzel (1987).

The distribution of these methods across the 13 countries is provided in Table 3.1.

**Table 3.1:** Methods of analysis employed by country

| Method | Countries |
|---|---|
| AuPA | Netherlands, USA |
| AcPA | Italy |
| AuPA+AcPA | Australia, Austria, Brazil, China, Germany, Netherlands, Spain, Turkey, UK, USA |
| HASR | Spain, Germany, South Korea, Sweden, USA |

The distribution of the methods of analysis relative to type of workplace is shown below in Table 3.2.

**Table 3.2:** Places of work against method of analysis employed

| | AuPA | AcPA | AuPA+AcPA | HASR |
|---|---|---|---|---|
| **university or research institute** | 2 | 1 | 13 | 3 |
| **government laboratory/agency** | | | 8 | 4 |
| **private laboratory** | 1 | | 7 | 1 |
| **as an individual** | | | 7 | |

As is evident in Table 3.2, the HASR method is used most frequently by government laboratories/agencies (33% use it), as opposed to only 16% using HASR in universities or research institutes. AuPA + AcPA is well distributed across all places of work.

The specific features of speech that are analyzed and considered important vary from analyst to analyst within each of the method categories. The data relating to this variation are presented in § 3.9.

## 3.5 Conclusion Frameworks

As with method of analysis, there is no consensus within the forensic speech science community as to how conclusions are and should be expressed.

Currently, there is much debate in the field about the 'logical' and 'legally correct' frameworks (French and Harrison, 2007; Rose and Morrison, 2009; French et al., 2010).

A variety of frameworks for expressing conclusions is currently utilized across the world. The conclusion frameworks may be grouped under the following headings:

***Binary Decision:***

A two-way choice that either the criminal and suspect are the same person or different people.

***Classical Probability Scale (CPS):***

The probability or likelihood of identity between the criminal and suspect is stated. Typically, the assessment is a verbal rather than a numerical one and it may use such terms as "likely/ very likely to be the same (or different) speakers." These types of judgments are often labelled as 'frequentist'.

Different probability scales are used by different experts, causing concern amongst practitioners over the lack of clarity caused by these different scales used for making conclusions (Broeders, 1999, p. 229). DNA evidence conclusions (presented using Likelihood Ratios) in combination with "the scientific status of forensic evidence" in the USA have had a large impact on all fields of forensics in "gradually undermining the traditional use of probability scales" (Broeders, 1999, p. 231). Conclusions made using CPSs do not (intentionally or otherwise) incorporate any estimate of typicality, and generally fail to acknowledge the defense hypothesis (e.g. the evidence came from/was produced by someone other than the suspect).

***Likelihood Ratio (LR):***

This expresses the results as the likelihood of finding the degree of correspondence or non-correspondence between the samples on the basis of the prosecution hypothesis (that they come from the same speaker), against the defense hypothesis (that they come from different speakers). Some analysts express the likelihood ratio as a number; others do so verbally. Both verbal and numerical LRs provide a strength of evidence statement (see § 2.3.1 for more information) in the form of a verbal or numerical conclusion, respectively (Morrison, 2009).

Using LRs, unlike CPSs, allows for typicality assessments to be made. This requires population statistics or a knowledge of the population in question for a given piece of evidence. In light of the "paradigm shift" (Morrison, 2009), LRs are thought to be the most "logical" way in which to express conclusions. Furthermore, the National Research Council (NRC) report to Congress on Strengthening Forensic Science in the United States recommends Aitken and Taroni (2004), Evett (1990), and Evett et al. (2000), as they provide "the essential building blocks for the proper assessment and communication of forensic findings" (2009, p. 186). All three are proponents of the likelihood ratio framework.

> ***UK Position Statement:***
>
> This conclusion framework potentially involves a two-part decision. The first part concerns the assessment of whether the samples are compatible, or "consistent", with having come from the same person. The second part, which only comes into play if there is a positive decision concerning consistency, involves an evaluation of how unusual or "distinctive" the features common to the samples may be (French and Harrison, 2007).

Expanded explanations of these various frameworks are to be found in, *inter alia*, Broeders (2001), Champod and Evett (2000), French and Harrison (2007), French et al. (2010), Jessen (2008), Morrison (2009c), and Rose and Morrison (2009).

Some methods of analysis lend themselves more readily than others to the adoption of certain conclusion frameworks. For example, some automatic systems express the results of the comparison as a numerical LR as one of their options. A breakdown of methods against conclusion frameworks appears in Table 3.3.

**Table 3.3:** Methods used for analysis in forensic speaker comparisons against conclusion frameworks

| | Binary Decision | CPS | Numerical LR | Verbal LR | UK Position Statement | Other |
|---|---|---|---|---|---|---|
| **AuPA** | | 1 | | | | 1 |
| **AcPA** | | | 1 | | | |
| **AuPA + AcPA** | 2 | 10 | 1 | 2 | 10 | |
| **HASR** | | 3 | 2 | 1 | 1 | |

As seen in Table 3.3, there is a tendency for participants using AuPA + AcPA to adopt the classical probability scale and UK Position Statement conclusion frameworks.

Table 3.4 breaks down conclusion frameworks by country. Some countries appear more than once, as there were multiple respondents from the same country, with individual experts implementing different conclusion frameworks.

**Table 3.4:** Conclusion frameworks used by country

| Conclusion Framework | Countries |
|---|---|
| Binary Decision | Brazil, China |
| Classical Probability Scale | Australia, Austria, Brazil, Germany, Netherlands, South Korea, Sweden, UK, USA |
| Numerical LR | Australia, Germany, Italy, Spain |
| Verbal LR | Netherlands, USA |
| UK Position Statement | Germany, Spain, Turkey, UK, USA |

A Likert Scale was used to measure the respondents' level of satisfaction with the conclusion method s/he used. Likert ratings were averaged across respondents. The scale ranged from 1 (extremely dissatisfied) to 6 (extremely satisfied). Table 3.5 reports the number of experts responding, mean scores, and standard deviations for satisfaction levels by conclusion frameworks.

**Table 3.5:** Satisfaction with conclusion framework

| Conclusion Framework | Mean Likert Rating | Standard Deviation | Number of Experts |
|---|---|---|---|
| Numerical LR | 5.00 | 0.00 | 4 |
| UK Position Statement | 4.27 | 0.65 | 11 |
| Verbal LR | 4.00 | 0.00 | 3 |
| Classical Probability Scale | 3.69 | 0.95 | 13 |
| Binary Decision | 3.50 | 2.12 | 2 |

### 3.5.1 Population Statistics

Out of all respondents, 70% reported that they use some form of population statistics in arriving at their conclusions. 58% stated that they had personally collected population statistics for the incidence of occurrence of one or more phonetic or acoustic features.

The features to which the statistics relate include fundamental frequency (used by almost all of the 70%), articulation rate, voice onset time, long term formant frequencies, and, where applicable, stammer/stutter. A number of the respondents commented that if more population statistics were available they would use them.

## 3.6 Guidelines

Respondents were asked whether they followed a protocol/set of guidelines in each forensic speaker comparison case and if so whether they had been involved in its design. 85% of respondents used some form of a protocol or set of guidelines. For those following a protocol or guidelines in casework, respondents were asked how their protocol/guidelines came into existence. The responses are distributed in Table 3.6.

**Table 3.6:** Creation of guidelines/protocols

| Origin | Number |
|---|---|
| Developed personally | 6 |
| Developed in conjunction with colleagues | 17 |
| Given it by place of work | 3 |

## 3.7 Casework Analysis: Alone or in Conjunction with Others

When asked whether they worked individually or in conjunction with colleagues in carrying out speaker comparisons, participants provided the information set out in Table 3.7.

**Table 3.7:** Workers involved in a single case

| How Work is Carried Out | Number |
|---|---|
| All work done individually | 13 |
| Work done with the help of an assistant | 4 |
| Work done in conjunction with other members of a group/team | 10 |
| Work is done individually and checked by someone else | 7 |

## 3.8 Casework in Foreign Languages

Besides carrying out casework in their native language, 56% of experts stated that they also conduct casework in other languages. Collectively, these experts have worked on cases in over 40 different languages other than their own.

Of those who work with other languages, 94% require the assistance of a native speaker of the language in question, and, of those requiring such assistance, 56% deem it necessary for the assistant to have a qualification in linguistics and/or phonetics.

## 3.9 Features Examined in Detail

This section reports on the aspects of recorded speech that respondents take into account or consider important in FSC cases.

### 3.9.1 Phonetic Features

Respondents were asked whether and with what frequency they examined the following features.

#### 3.9.1.1 Segmental Features

All respondents analyze vowel and consonant sounds in the course of their examinations. With regard to **vowels**, 81% invariably carried out some form of analysis and 13% routinely did so. 94% of all experts evaluated the auditory quality of vowels, 97% carried out some form of formant examinations and 58% measured vowel durations.

Of those undertaking **formant** examinations, all measure the second resonance (F2). 87% of respondents reported measuring F1 and an equal percentage reported measuring F3. 17% of respondents stated that they measure F4. Only 10% of respondents measuring formants measured F1-F4, 63% measured F1-F3, and 10% measured either F1 and F2 or F2 and F3. In respect of which **aspects of formants** are examined, 94% reported measuring center (i.e. temporal midpoint) frequencies of formants of monophthongs, 71% reported measuring formant trajectories of diphthongs and 45% examined vowel-consonant or consonant-vowel formant transitions. 35% stated that they examine formant bandwidth and 13% reported examining formant densities.

In relation to **consonants,** all respondents reported subjecting them to some form of examination; 52% invariably did so. For all experts, 88% of respondents reported evaluating auditory quality. 82% stated that they examined aspects of timing and 48% reported measuring the frequencies of energy loci.

Table 6 reports the frequency with which consonants, broken down by manner of articulation, are analyzed in FSC cases. Respondents gave their answers using a 6-point Likert Scale ranging from 1 (never) to 6 (always). The number of experts responding, mean Likert ratings, and standard deviations are represented in Table 3.8 for those respondents who are native English speakers (and working on English cases only).

**Table 3.8:** Frequency of consonant analysis in English

| Manner of Articulation | Mean Likert Rating | Standard Deviation | Number of Experts |
|---|---|---|---|
| Fricatives | 4.85 | 1.21 | 13 |
| Plosives | 4.73 | 1.49 | 11 |
| Approximants | 4.50 | 1.27 | 10 |
| Laterals | 4.46 | 1.13 | 13 |
| Nasals | 4.08 | 1.24 | 12 |
| Affricates | 3.82 | 1.47 | 11 |
| Taps/Flaps | 3.70 | 1.77 | 10 |
| Trills | 3.18 | 2.04 | 11 |

#### 3.9.1.2 Suprasegmental Features

All respondents (excluding those using AuPA only) routinely measure **fundamental frequency** in their comparisons. With respect to what they measure, for those conducting some form of AcPA, 94% reported measuring the mean, 41% the median, 34% the mode, 72% standard deviation, 25% the alternative baseline (the value (in Hz) that falls 7.64% below the F0; Lindh, 2007), and 6% measure the range. Considering all aspects of fundamental frequency listed above, the most common combination for analysis was the mean plus the standard deviation (22% of participants), followed by 19% examining mean, median, mode and standard deviation together. Only 9% measure the mean, median, mode, standard deviation, and alternative baseline. Single respondents also reported measuring the coefficient of variation, also known as the 'varco' (the standard deviation divided by the mean (Jessen et al., 2005)), the first and third quartiles, and kurtosis/skew. It is important to note that although many respondents reported analyzing the fundamental frequency, a large proportion point out that it is usually of little help. One respondent stated that, "[fundamental frequency is] usually used as an elimination tool rather than an identification tool."

94% of the respondents who include an AuPA stated that they examine **voice quality** as part of their overall procedure, although only 77% of these invariably or routinely examine it. Further to this, 61% of those who examine voice quality do so using a recognized scheme (e.g. Laver, 1980) or modified version of such a scheme, for its description. Of those experts examining voice quality the large majority (63%) reported using the Laver Voice Profile Analysis Scheme (VPAS) or a modified version of it. 21% of experts perform an auditory analysis of voice quality and provide some form of a verbal description. The remaining experts use the GRBAS scheme (Grade, Roughness, Breathiness, Asthenia, Strain; see Bhuta et al. 2004 for more information) or a modified version of it (13%), and a single expert (3%) reported using LTS spectra (long-term spectra) for examining voice quality. Furthermore, three experts provided insightful commentary regarding the discriminant power of voice quality, "[voice quality] can often be central to the analysis and is best analyzed systematically using a detailed scheme such as the Edinburgh VPA." Another respondent states that "voice quality is frequently strongly discriminating [in forensic speaker comparisons]," and the third expert comments that they are "increasingly of the view that voice quality is one of the most valuable but least well understood" parameters.

85% of all respondents stated that they examine **intonation** with one or another level of frequency. However, of these only 25% look at intonation invariably. The specific aspects of intonation vary, with tonality[11] being reported more than tonicity[12], 67% vs. 38% of respondents (Ladd, 1996, p. 10). Tails of tone units were examined by 46% and heads by 29%.

93% of respondents stated that they analyze **tempo** with varying degrees of frequency. Of those analyzing tempo, 81% apply a formal measure (e.g. speaking rate (SR) or articulation rate; Künzel, 1997). For formal measures, articulation rate (AR) was reported most frequently by 47% of respondents compared to only 19% that use speaking rate and 16% that use both articulation rate and speaking rate. Those using AR were asked how they defined a syllable, and 93% of the respondents reported using phonetic syllables for AR rather than linguistic ones. 73% stated that they examine speech **rhythm** with varying regularity.

## 3.9.2 Non-Phonetic Features

### 3.9.2.1 Higher-Order Linguistic Features

In addition to examining phonetic features, 76% of all respondents reported examining **discourse features** and/or **conversational behaviors** (discourse markers, aspects of turn-taking, telephone opening and closing behaviors, patterns of code switching). 88% of all experts stated that they examine **lexico-grammatical** usage. Lexical features were examined most frequently, followed by syntax and morphology.

[11] "Tonality marks one kind of unit of language activity, and roughly where each such unit begins and ends: one tone group is as it were one move in a speech act" (Halliday, 1967, p. 30).

[12] "Tonicity marks the focal point of each such unit of activity: every move has one (major), or one major and one minor, concentration point, shown by the location of the tonic syllable, the start of the tonic" (Halliday, 1967, p. 30).

#### 3.9.2.2 Non-Linguistic Features

94% of the respondents who answered this question reported examining non-linguistic features at least some of the time. In descending frequency order, specific features were as follows: filled pauses, tongue clicking[13], audible breathing, throat clearing, and laughter.

## 3.10 What is Considered Discriminant

In addition to being asked about features within the linguistic, phonetic and acoustic domains, respectively, participants were given the opportunity to identify which feature from *any* domain they found most useful for discriminating speakers. For all respondents together, voice quality was reported most often (32%), followed by dialect/accent variants and vowel formants (both 28%). 20% reported speaking tempo and fundamental frequency as useful parameters. This was followed by rhythm (16%). Lexical and grammatical choices, vowel and consonant realizations, phonological processes (e.g. connected speech processes) and fluency were all reported by 13% of the respondents. One respondent went as far as stating that vowel formant analysis "is rarely insightful."

Interestingly, though perhaps not surprisingly, the vast majority of participants alluded to the fact that despite some individual parameters having significant weight, it is the overall *combination* of features that they consider crucial in discriminating between speakers. In Aristotelian terms, 'The whole is greater than the sum of the parts" (Aristotle, Metaphysica 10f-1045a).

[13] It is perhaps better to classify tongue clicking as a linguistic feature when it is used in a inherently functional way, such as that described in Chapter 7. However, at the time of the survey, clicks were classified as non-linguistic.

## 3.11 Discussion

The purpose of this chapter has not been to advance any argument or to develop theoretical propositions. Rather, its objective has been the much more mundane one answering to the motivations set out in §3.1: laying out basic factual information concerning the practice of FSC internationally in the present day and drawing upon the collective expertise of FSC experts worldwide so as to identify current working methods and features of speech that are considered to have the greatest potential for discriminating between individuals.

Those not directly involved in this specialist field but working, for example, in other aspects of phonetics or linguistics, may well be surprised at the lack of consensus over such fundamental matters as how speech samples are to be analyzed and compared, which aspects of the samples are to be assigned greatest importance during the analytic process, and how conclusions are to be expressed at the end of it. However, we are assured by those working in various other fields of forensic science that the level of dissensus uncovered by the present survey is by no means unique to forensic speaker comparison. Indeed, some of the practices and preferences found here are undoubtedly dictated or constrained by the rules of the institutions and firms in which the participants work. Where those organizations include other forensic science disciplines, the options, particularly for the framing of conclusions, may be laid down unilaterally for all types of casework investigation undertaken under their auspices. For instance, the Dutch government forensic science facility, the Netherlands Forensic Institute, requires that the outcomes of every investigation undertaken by its employees, irrespective of the forensic discipline, be expressed within a Bayesian likelihood ratio framework (Meuwly,

p. c.). Likewise, one of the participants in the present study who used a binary decision format when expressing conclusions stated that to do so was a requirement of his/her employer.

Over and above the rules laid down by public and private sector laboratories, nations may also set down requirements, either by statute or via case law. Some jurisdictions, notably the England and Wales division of the UK, have been extremely non-prescriptive in this respect, according the expert a very high degree of autonomy and discretion over the methods of analysis he/she adopts and the way the outcomes are formulated. In respect of forensic speaker comparison evidence, the England and Wales position was affirmed in the Appeal Court ruling R -v- Robb (1991), in which the court ruled that whether or not an expert used any acoustic testing was entirely his/her own decision, and re-iterated in relation to the same issue in the more recent appeal R -v- Flynn and St John (2008). Indeed, the main analytic issues over which the higher courts have seen fit to pass down general prohibitions to forensic experts concern the use of statistics in representing the strength of evidence (cf. R -v- Doheny and Adams (1996); R -v- T (2011)). Where experts enjoy freedom of choice, one might expect their preferences to be influenced by individual intellectual commitments. However, in spite of the latitude allowed by the UK legal system, it is of note that all nine UK experts taking part in the study use the combined AuPA + AcPA method. Indeed, this method is the predominant one across all countries represented in the survey (25 = AuPA + AcPA; 10 = other – see Table 3.3). Thus, although the results show a wide range of variation in methods, there is nevertheless a very large degree of convergence.

As for the future, certain trends can be predicted. One is that as time goes on and further improvements are made to the error rates of (semi)automatic systems and to their capabilities for handling real case (i.e., non-studio) recordings, one would expect to find such systems increasingly being incorporated into casework alongside the AuPA + AcPA approach. This development would be particularly apposite in the USA, where the appeal court ruling Daubert -v- Merrell Dow Pharmaceuticals Inc. (1993) is taken by many lower courts as the benchmark for admissibility of expert evidence, and within that ruling is the statement that 'the court ordinarily should consider the known or potential error rate' of the method. ASRs readily lend themselves to meeting this criterion, and, indeed, many systems are subject to such testing as part of the annual NIST evaluations (Greenberg et al., 2010). Further, a number of ASRs have an LR as one of their easily selectable options for representing the results of speaker comparisons. As seen in Table 3.3, most experts currently express their conclusions in terms of a classical probability scale (14), whilst only half as many (7) use some form of LR (3 = verbal LR; 4 = numerical LR). The increasing use of ASR software, together with the current 'paradigm shift' in forensic science towards Bayesian reasoning, and the use of LRs for presenting results, would lead one to expect an increase in the number of experts using LRs and a corresponding decrease in the use of other conclusion frameworks.

Morrison (2009c, p. 298) suggests that "today we are in the midst of what Saks and Koehler (2005) have called a *paradigm shift* in the evaluation and presentation of evidence in the forensic sciences which deal with the comparison of the quantifiable properties of objects of known and questioned origin, e.g., DNA profiles, finger marks, hairs, fibres, glass fragments, tool marks, handwriting, and voice recordings." However, he fails to acknowledge the fact that not all speech evidence is of the quantifiable type, as demonstrated in the survey results from the present study. The Bayesian framework of likelihood ratios has been adopted by many fields in the forensic disciplines where quantifiable evidence is of the norm and qualitative evidence is something that does not necessarily come into question (e.g. DNA or fingerprints). It is important to recognize that speech does not consist entirely of measurements. There are elements of speech that are best described/analyzed qualitatively (i.e. certain aspects of voice quality (e.g. lingual body orientation), lexical, syntactic, or morphological choices, audible breathing, laughter). If such features can be quantified in some form, then it is plausible that we will one day see an entire forensic speaker comparison case completed in a Bayesian framework, but until then there will still be experts who will continue to present such features in a qualitative form, whether that is alongside a LR conclusion or another form of conclusion (e.g. CPS or UK Position Statement).

Additionally, in light of the popularity of the Bayesian framework, it can be predicted that more research on LRs will be carried out. This can be seen as a positive trend, as parameters that experts found to be discriminant in their experience (as reported in this survey), may now be tested empirically, and general strength of evidence statements can potentially be attributed to certain features. Given this, experts and researchers in the field of forensic speech science can give appropriate weight in forensic casework to those features found to be most discriminant through intrinsic[14] and extrinsic[15] likelihood ratio testing.

Finally, those differences that currently exist across practitioners may be reduced through blueprints[16] and drives for international co-operation and cross-border transferability of forensic science evidence (e.g. House of Commons Northern Ireland Affairs Committee, 2009). And, of course, the prerequisite for resolution of differences is knowledge of their existence. Insofar as the present study lays bare that information, it may be considered to be making a modest first step towards international unity.

## 3.12 Limitations

The International Survey on Forensic Speaker Comparison Practices has three limitations. The first is the limited number of experts who took part in the survey. Ideally, one would like to work with a larger sample size in order to represent the total population of forensic speech scientists as accurately as possible. Thirty-six is a large proportion of practicing forensic phoneticians. However, it would have been preferable to include even more forensic phoneticians and to have been able to represent a greater number of countries, languages, and methods in order to achieve the most accurate representation of current practices in forensic speaker comparison.

The second limitation is the lack of representation from those experts using ASR alone. As is evident in the NIST speaker recognition evaluations

[14] Intrinsic likelihood ratio testing uses the same set of speakers (e.g. from the same speech corpus) for both the test and reference samples.

[15] Extrinsic likelihood ratio testing uses different sets of speakers (e.g. from different speech corpora) for the test and reference samples.

[16] Blueprint is used here to refer to some form of standards documents.

(Campbell et al., 2009), there are a number of experts around the world who use ASR alone, and the survey results presented in this chapter fail to represent this fact. Despite efforts to recruit participants that utilize ASR alone, no such experts responded to the survey.

The final limitation is the simple fact that these results have a 'limited shelf life', meaning that the field is always changing and forever evolving, and these results are only a snapshot of the field as it stood in 2010-2011. The trends seen in the survey will certainly vary in the future as more research is carried out and new methodologies are put into practice.

## 3.13 Parameters Chosen for Further Analysis

As stated in § 1.3, this survey served in part as a hunting ground for identifying the speech parameters believed by experts to hold the greatest discriminant potential. Based on responses from the practitioners, I now identify four parameters from the survey that experts found to be highly discriminant and/or analyzed relatively often in casework: articulation rate, long-term formant distributions, long-term fundamental frequency, and clicks (velaric ingressive stops). The subsequent chapters analyze each of the four parameters in turn, while referring to the discriminant expectations of a given parameter.

# Chapter 4: Articulation Rate

## 4.1 Introduction

Forensic phoneticians have suggested that speech tempo is an important parameter for forensic speaker comparisons, with 93% of experts analyzing speech tempo and 73% of those doing so with varying regularity (Chapter 3). It is also reported that when asked which parameters they found highly discriminant in forensic speaker comparisons, 20% of all experts reported that they found speech tempo to be the most useful parameter for discriminating speakers. Overall, speech tempo was ranked as the third most helpful parameter (alongside F0) of all possible parameters used in a forensic speaker comparison. Analyzing speech tempo in detail for a large, homogeneous group of individuals provides insight into the distribution of and variation within the parameter and thereby its ability to discriminate between speakers.

In forensic phonetics, speech tempo is typically quantified as either speaking rate (SR) or articulation rate (AR; Künzel, 1997). Both speaking and articulation rate measure the speech tempo of an individual, but the two measures capture slightly different aspects of tempo. Speaking rate (SR) can be defined as “the rate of speech of the whole speaking-turn. It therefore includes all speech material (linguistic or non-linguistic), together with any silent pauses, that are contained within the overall speaking-turn” (Laver, 1994, p. 158). Articulation rate (AR) is defined as “the rate at which a given utterance is produced. The speech material measured by articulation rate therefore excludes silent pauses by virtue of the definition of an utterance, which begins and ends with silence” (Laver, 1994, p. 158). The difference between speaking rate and articulation rate is that the former includes disfluencies and filled/unfilled pauses in the calculation, whereas the latter excludes disfluencies and unfilled pauses. Within the field of forensic speech science the majority of experts report a preference for measuring articulation rate rather than speaking rate in forensic speaker comparison casework (Chapter 3).

Population statistics for articulation rate on a large, homogeneous scale (100+ speakers) exist for German and Chinese, but as yet, there has not been a similar study carried out on English. This study presents the analysis for the ARs and standard deviations (SD) of 100 Southern Standard British English (SSBE) male speakers. The results concern both the inter-speaker and intra-speaker variation of AR, as well as assessing the evidential value of AR as a parameter in forensic speaker comparisons.

## 4.2 Literature Review

Articulation rate has been investigated in British English in small-scale studies. Goldman-Eisler (1956) was one of the first to analyze and calculate articulation rate in a population. In her study, she examined the spontaneous speech of eight British adults in 30- to 60-minute interview-type recordings. AR was calculated by counting the number of syllables (the definition of the syllable and interval type were undefined in the study) in an utterance, with an utterance defined as "periods of speech lasting from a preceding question or utterance of an interviewer to the next, which is usually occasioned by the subject having come to a natural stop or pause" (Goldman-Eisler, 1956, p. 137). It was found that the mean AR across speakers was 4.95 syllables per second, ranging from 4.4 for the slowest to 5.9 for the fastest speakers. The mean standard deviation across speakers was 0.91 syllables per second (syll/sec), ranging from 0.54 to 1.48.

Kirchhübel and Howard (2011) also collected articulation rate figures for British English, while investigating properties of speech that could potentially be correlated with emotional/psychological stress. The study examines the spontaneous speech of a loosely homogeneous group of 10 young Southern British males in mock police interviews. Along with the AR and SR results for the psychologically-stressed speech of the subjects, Kirchhübel and Howard (2011) also provided baseline results for the speakers using interpause stretches to obtain AR measurements. The mean AR for speakers was 5.81 syll/sec with a range of 5.14 to 7.00 and a standard deviation of 0.89 syll/sec (range 0.79 to 1.01).

In a later study, Goldman-Eisler (1968) further examined AR as well as SR. However, this study focused on intra-speaker variation. It was observed that AR exhibits fairly little intra-speaker variation, whereas there is much more variability present in SR. Henze (1953) conducted a similar investigation in German using spontaneous speech in the form of story-telling elicited by a film, and the same observations were made. It is noted that different speech tasks, for example read versus spontaneous speech produced in different emotional states, can cause differences in the pauses that speakers use (e.g. number, kind, duration), in turn causing variations in speech rate across different speaking tasks. Articulation rate differed slightly across the different tasks, but it was found to be relatively stable across tasks in these two studies.

With respect to the implications for forensic speaker comparison, Künzel (1997) examined AR, SR, and various pausing parameters in German. He retested claims that intra-speaker variation was lower in AR than SR. He confirmed prior results showing that intra-speaker variation is much smaller. For the experiment, the read and spontaneous speech of five males and five females was analyzed, and SR was found to be higher in read speech than in spontaneous speech. This is largely due to the fact that speakers use far fewer hesitation pauses (i.e. filled and unfilled) in read speech than in spontaneous. AR, on the other hand, was not significantly different between read and spontaneous speech, and AR for individual speakers had coefficients of variance that were smaller than they were with SR. To further evaluate the possible discriminating power of SR and AR, Künzel looked at cumulative distributions of both intra- and inter-speaker differences. According to the equal error rates (see § 4.6.1) calculated, AR was found to have more speaker-discriminating power than SR.

Following Künzel's (1997) conclusion that AR is a better discriminator than SR, investigators have begun to examine AR in more detail. In keeping with Künzel's conclusion that "AR will have to be interpreted with caution when used in forensic speaker recognition until its possibilities and limitations have been assessed on the basis of genuine case material and large numbers of speakers" (1997, p. 79), additional studies have been conducted in both German and Chinese. Jessen (2007b) analyzed the AR of 100 male speakers of German. AR was measured for both spontaneous and read speech. It was found that, contra Künzel (1997), the mean AR was significantly higher in read speech. Overall, Jessen found the mean AR for the 100 speakers was 5.21 syll/sec. In order to calculate ARs, he was the first to implement a new methodology in which "memory stretches" (Jessen, 2007b, p. 53) were utilized rather than "interpause stretches" and "intonation phrases" (Trouvain, 2004, p. 50), which had been commonly used in previous studies. Jessen describes the methodology behind "memory stretches" as "the phonetic expert [going] through the speech signal and [selecting] portions of fluent speech containing a number of syllables that can easily be retained in short-term memory." After listening several times the expert then counts the number of syllables that he/she is able to recall from memory to be included in this portion of speech (Jessen, 2007b, pp. 54-55).

Cao and Wang (2011) followed the methodology of Jessen (2007b) and examined the ARs for 101 male Mandarin Chinese speakers in spontaneous telephone speech. They investigated inter-speaker and intra-speaker variation in AR, and found both the global ARs (GAR) and means of local ARs (LARmean) to be fairly normally distributed (GAR and LAR mean are explained further in § 4.3.2). The mean global articulation rate (GAR) was 6.58 syll/sec and the mean of the local articulation rates (LARmean) was 6.66 syll/sec. They also report that the range of AR for a given speaker is relatively small and stable. However, ARs in Mandarin Chinese appeared to be higher than English and German studies. The authors attribute the difference to the simpler syllable structure found in Chinese. Chinese syllables are largely /CV/ in shape; therefore more syllables per second can be produced than is possible with the inherently longer syllables in German and English (Cao and Wang, 2011, p. 398).

Although AR has been examined in large-scale studies of both German and Chinese, the greatest number of subjects examined in a previous study of English speakers is 50, and many studies examined are based on considerably fewer. Table 4.1 provides an overview of AR studies conducted on English.

**Table 4.1:** Overview of articulation rate studies

| Study | Subjects | Task | AR mean avg. | AR range | SD mean avg. | SD range |
|---|---|---|---|---|---|---|
| Goldman-Eisler (1956) | British: 8 subjects | Interviews: Spontaneous | 4.95 | 4.4-5.9 | 0.91 | 0.54-1.48 |
| Robb et al. (2004) | American: 20 male and 20 female adults | Rainbow Passage: Read | 5.27 |  | 0.40 |  |
|  | New Zealand: 20 male and 20 female adults | Rainbow Passage: Read | 5.70 |  | 0.47 |  |
| Doherty & Lee (2009) | Irish: 22 males | Read (1st time through Rainbow Passage) | 5.68 |  |  |  |
|  | Irish: 22 males | Read (2nd time through Rainbow Passage) | 6.05 |  |  |  |
|  | Irish: 22 males | Conversation: Spontaneous | 5.88 |  |  |  |
|  | Irish: 22 females | Read (1st time through Rainbow Passage) | 5.38 |  |  |  |
|  | Irish: 22 females | Read (2nd time through Rainbow Passage) | 5.67 |  |  |  |
|  | Irish: 22 females | Conversation: Spontaneous | 5.58 |  |  |  |
| Jacewicz et al. (2009) | North Carolina: 50 adults | Conversation: Spontaneous | 5.41 |  | 0.48 |  |
|  | North Carolina: 50 adults | Sentences: Read | 3.27 |  | 0.44 |  |
|  | Wisconsin: 44 adults | Conversation: Spontaneous | 4.81 |  | 0.54 |  |
|  | Wisconsin: 44 adults | Sentences: Read | 3.54 |  | 0.34 |  |
| Kirchhübel & Howard (2011) | SSBE: 10 males | Interviews: Spontaneous | 5.81 | 5.14-7.00 | 0.89 | 0.79-1.01 |

As can be seen in Table 4.1, a number of studies have examined AR for both read and spontaneous speech, but to date only two small-scale studies on British English have been carried out (Goldman-Eisler, 1956; Kirchhübel and Howard, 2011). Combined, the results for ARs from these in respect of spontaneous speech have a mean rate of 5.29 syllables per second, with the slowest mean at 4.81 syll/sec and the fastest at 5.88 syll/sec. It is important to note that these figures are the result of studies of only a few varieties of English, and how other varieties and dialects may pattern is unknown. The most recent AR study on British English (Kirchhübel and Howard, 2011) has a difference of more than 1.00 syll/sec relative to Goldman-Eisler's (1956) study carried out about 55 years earlier. It is hypothesized that the results of the present study will pattern more closely with those of Kirchhübel and Howard (2011) than those of Goldman-Eisler (1956), as the former is based on a more demographically and linguistically homogeneous group of speakers and uses a similar methodology to the present study (the methodology in Goldman-Eisler (1956) is not transparent and is therefore difficult to compare).

## 4.3 Population Statistics for Articulation Rate

The following section presents the collection of population statistics for articulation rate in a large, homogeneous group of 100 male speakers. This data serves as the first of its kind in providing detailed information on the distribution of and variation in articulation rate for a large group of individuals who speak SSBE.

### 4.3.1 Data

The data for the current chapter as well as subsequent chapters come from the Dynamic Variability in Speech database (DyViS) recorded at the University of Cambridge (Nolan et al., 2009). The DyViS database is a large speech corpus collected under simulated forensic conditions (de Jong et al. 2007). It is comprised of recordings of 100 male speakers of Southern Standard British English (hereafter referred to as SSBE) aged 18-25. This group of speakers is meant to represent a homogeneous population in respect of sex, age, and accent group. All speakers were recorded under both studio and telephone recording conditions for Task 2 (see below), and under studio recording conditions for a number of different speaking styles (i.e. Task 1, Task 3, and Task 4). The DyViS recordings include four tasks identified below (adapted from de Jong et al., 2007):

Task 1: simulated police interview (studio quality)
Task 2: telephone conversation with 'accomplice' (studio and telephone quality)
Task 3: reading passage (studio quality)
Task 4: reading sentences (studio quality)

The first task in DyViS is a simulated police interview, whereby the speaker is interrogated in a mock police investigation in relation to a (fictional) drug trafficking crime. The speech is spontaneous insofar as speakers were given visual stimuli (e.g. pictures of people and places) to prompt the construction of their responses to the investigator (interlocutor). There were a number of target words from the visual stimuli that were elicited by the interlocutor (i.e. the interlocutor asked the speaker specific questions in order to elicit the target words). The second task in DyViS is a telephone conversation between the speaker and his accomplice, 'Robert Freeman' (the interlocutor is the same person for all 100 speakers). Task 2 is recorded from the studio end of the telephone call as well as via an intercepted external BT landline. The second task, like Task 1, is spontaneous speech, whereby the interlocutor questions the speaker in a mock police interview. The interlocutor for Task 2 elicits from the speaker the same target words as those used in the police interview, which inevitably leads the discussion in Task 1 and Task 2 to be very similar.

Tasks 3 and 4 of DyViS are both forms of read speech. Task 3 consists of a read news report pertaining to the alleged drug trafficking crime. The same target words are included in the read report. Task 4 is read speech from controlled sentences that have a large number of SSBE vowels in nuclear non-final position (i.e. in closed syllables), with six repetitions each.

The studies carried out in the remainder of this thesis will include only data from either Task 1 or Task 2 (studio quality, spontaneous speech) of the DyViS database. The current chapter uses Task 2 for calculating articulation rate.

The DyViS studio recordings were all made using a Marantz PMD670 portable solid state recorder at a sampling rate of 44.1 kHz and 16 bit depth (de Jong et al., 2007). All speakers were recorded via a Sennheiser ME64-K6 cardioid condenser microphone positioned approximately 20 cm from the speaker's mouth. The recordings were made in a sound-treated room in the Phonetics Laboratory at the University of Cambridge (Nolan et al. 2009:40).

### 4.3.2 Methodology

The Task 2 recordings used for the current study were 15 to 25 minutes in duration. However, only the relevant amount of material was analyzed in order to extract between 26 and 32 “memory stretches.” 26 was chosen as the lower boundary for “memory stretches”, because it was the maximum number of tokens that could be extracted from the shortest of the 100 recordings. The upper boundary of 32 was chosen semi-arbitrarily (only to have a large number of tokens for calculating likelihood ratios).

The general methodology employed in this study follows very closely that of Jessen (2007b). In measuring AR a number of decisions have to be made (Künzel, 1997; Trouvain, 2004). As Jessen (2007b, p. 53) explains, the first concern is the “kind of linguistic unit on the basis of which AR is counted.” As noted in Chapter 3, the majority of forensic phoneticians use the syllable as a unit of measurement, rather than sound segments or words, in turn producing AR rates in syll/sec as opposed to words per second (or minute). As a native speaker of a language, one has a fairly reliable intuitive ability to count the number of syllables in a specific stretch of speech. In terms of analysis, this avoids the need to become involved in examining intensity peaks in the acoustic signal on a syllable-by-syllable basis. Jessen also mentions that the syllable is “probably more a cognitive/linguistic unit grounded in the physics of speech production” (Keating, 1988, cited in Jessen, 2007b, p. 53). For these reasons, syllables in this study were determined auditorily through careful listening.

The second important decision for the measurement of AR relates to whether one should define syllables phonologically or phonetically. A phonological syllable is “defined in terms of the lexicon and grammatical rules of the language", whereas a phonetic syllable is one that is "manifested in phonetic reality" (Jessen, 2007b, p. 53). Jessen gives an example using the phrase "did you eat yet?" Phonologically we would count this as having four syllables; however, in reality the number of syllables may be reduced or in some rare cases even increased. If the phrase were to be reduced it might be realized as perhaps two syllables as in "jeet yet" (Jessen, 2007b). For this reason, the use of phonetic versus phonological syllables makes a difference in terms of AR counts. In a case where a phrase is phonetically only two syllables, AR will obviously be lower than if the same phrase was counted as four phonological syllables (see Jessen, 2007b, pp. 53-54 for further discussion). Jessen (2007b) suggests that syllables are best defined phonetically, rather than phonologically. This is because often in casework, speaker comparisons include speech from different dialects or foreign accents, and in certain cases it might be difficult to determine what the phonological form should be (Jessen, 2007b). However, Jessen (2007b) also notes that:

> [C]ounting actual syllables can lead to curious artefacts when a speaker in speaking rapidly deletes [phonetic] syllables, whereas another speaker might reduce or delete perhaps the same number of [phonetic] sounds but still preserves the number of underlying syllables. In such a case the former speaker ends up with lower AR than the latter although both would be about equally fast if AR were based on canonical rather than actual syllables.

Given that the present study is based on recordings of a linguistically homogeneous population with the same accent and that counting syllables phonetically has been shown to cause "curious artefacts[17]" (Jessen, 2007b, p. 54; Koreman, 2006), the current study is based on phonological syllables rather than phonetic syllables.

The third methodological decision, and the one which is perhaps the most influential on the results, involves the kind of speech interval that is selected for determining AR. The AR can be calculated for the entire duration of fluent portions in a recording. This number is known as "global AR". Alternatively, by taking multiple fluent speech stretches within a recording, "local ARs" can be calculated (Jessen, 2007b, p. 54). Miller et al. (1984) showed that speakers often change their speech tempo over the course of longer utterances. Therefore, in order to capture such changes in tempo that may occur within a single recording, it is more useful to obtain local ARs. Previously, researchers have commonly used interpause stretches and intonation phrases to identify speech intervals over which to calculate local ARs (Trouvain, 2004). Following Jessen (2007b), in order to avoid possible empirical or methodological problems associated with the two aforementioned methods of selecting speech intervals, a much simpler and more pragmatic approach was chosen for this study. Interpause stretches tend to result in intervals that are extremely variable in length due to pausing behaviors (which might reintroduce the influence of pausing which AR tries to eliminate; Jessen, 2007b). Intonation pauses are reliant on phonetic and linguistic judgments made by the analyst, which result in variation of the interval lengths depending on the expert's interpretation (Jessen, 2007b). Therefore, the speech interval used in the

[17] These 'artefacts' occur when one speaker may be speaking quickly and as a consequence deletes phonetic syllables, whereas another speaker is typically inclined to reduce or completely delete the same number of phonetic sounds. However, this speaker is able to preserve the number of underlying phonological syllables.

current study for computing local ARs is referred to as a “memory stretch” (Jessen, 2007b, p. 54). After listening several times to that interval of speech, the expert then counts the number of syllables that he/she is able to recall from memory being included in this portion of speech. Three examples of memory stretches from Speaker 036 are presented in the Table 4.2.

**Table 4.2:** Examples of memory stretches for Speaker 036

| | Memory Stretch | Number of Syllables | Time (in seconds) |
|---|---|---|---|
| (a) | I defended you gallantly | 8 | 0.984 |
| (b) | They wanted to know about a car park | 10 | 1.367 |
| (c) | They didn’t elaborate or anything | 11 | 1.161 |

Sony Sound Forge Audio Studio 10.0 was used for analysis. Speech intervals were only selected at least two minutes into the recording, to allow the speaker to become comfortable speaking to his accomplice in the presence of the recording equipment. Like Jessen, speech intervals containing fluent speech were chosen, and the region marked out. After listening several times, I would type out the speech phrase on the region marker tag. Following this, I would count the number of syllables included in that interval. After collecting enough memory stretches, it was possible to view all recorded regions that listed the number of syllables and included the length of the speech segment. Those figures were entered into Microsoft Excel and local and global ARs as well as standard deviations were computed for all speakers.

The procedure described above was applied to all 100 recordings analyzed in this study. The mean and standard deviation of AR for each speaker are used for analysis and are reported in § 4.3.2 below. The maximum number of syllables in a memory stretch was 26, but most stretches contained between

7 and 11 syllables (in order not to "push the limits[18]" (Jessen, 2007b, p. 55), and avoid mistakes following Jessen (2007b)). In line with the methodology implemented by Jessen, four syllables or more per stretch were used. The threshold is in place in order to avoid the "inclusion of very short interpause stretches that could unduly increase the effect of phrase-final lengthening on the calculated articulation rate" (Jessen, 2007b, p. 55). It is important to note that each memory stretch consisted of only fluent speech (for speech intervals). Fluent speech was defined as speech that did not include the following: any kind of pauses, either filled or unfilled, repeated syllables, unintelligible speech, and any syllable lengthening (judged subjectively) that went beyond canonical non-hesitation durations in English. The mean number of memory stretches measured per speaker was approximately 30, with a standard deviation of 2.1, a range of 26-32, and 2,993 total ARs calculated for the 100 speakers.

### 4.3.3 Results

The distributions of the local AR means and the standard deviations for individuals are presented in Figures 4.1 and 4.2. The y-axis represents the number of speakers that fall within a given range and the x-axis depicts articulation rate presented as syllables per second.

[18] That is in order to avoid trying to remember such an extensive interval of speech that mistakes are made when trying to recall it, as this could potentially affect the resulting ARs.

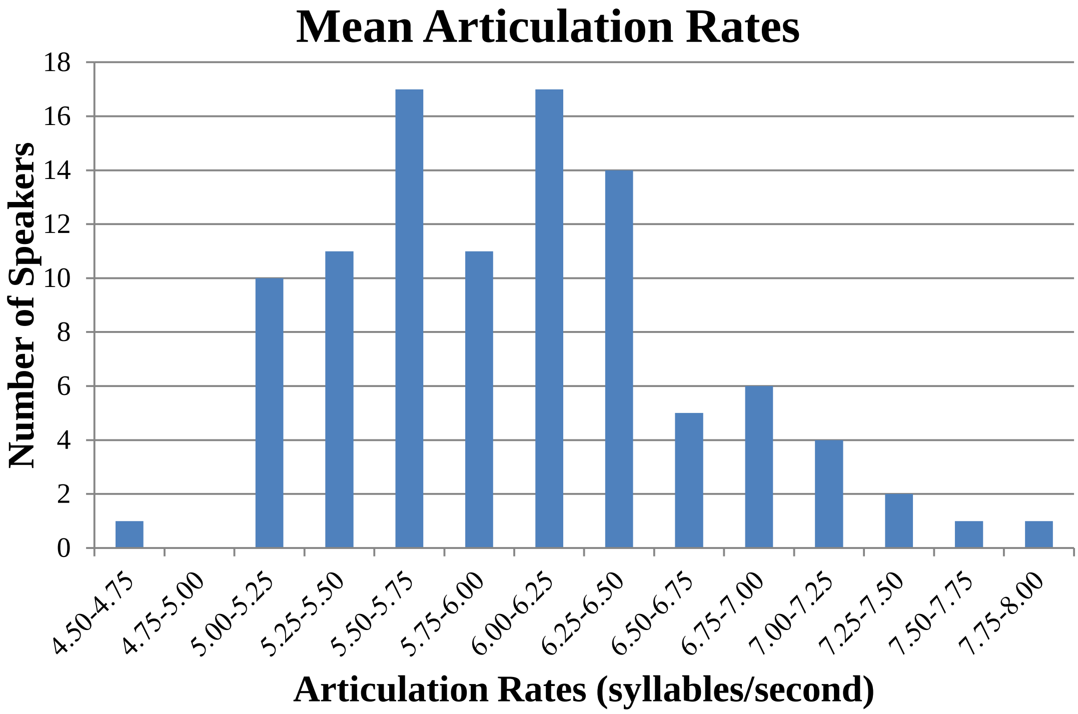


**Figure 4.1:** Distribution of mean articulation rates

There is a roughly normal distribution[19] for the mean ARs, as illustrated in Figure 4.1. The mean AR for the population is 6.02 syll/sec, with an overall range of mean AR from 4.57 to 7.79 syll/sec. The standard deviation of the mean is 0.64 syll/sec. The 100 speakers have mean ARs within a 3.22 syll/sec window.

The data were checked for two levels of outliers. This thesis defines suspected outliers as falling between 1.5 times the interquartile range (IQR) and 3 times the IQR, plus or minus the first or third quartiles. Any outliers that fall outside the upper bounds of 3 times the IQR are confirmed as definite or extreme outliers in this thesis. The mean AR has six suspected outliers at 7.23

[19] Normality was judged visually, and not through statistical testing.

syll/sec, 7.24 syll/sec, 7.47 syll/sec (x2), 7.53 syll/sec, and 7.79 syll/sec. However, there were no extreme outliers for mean AR.

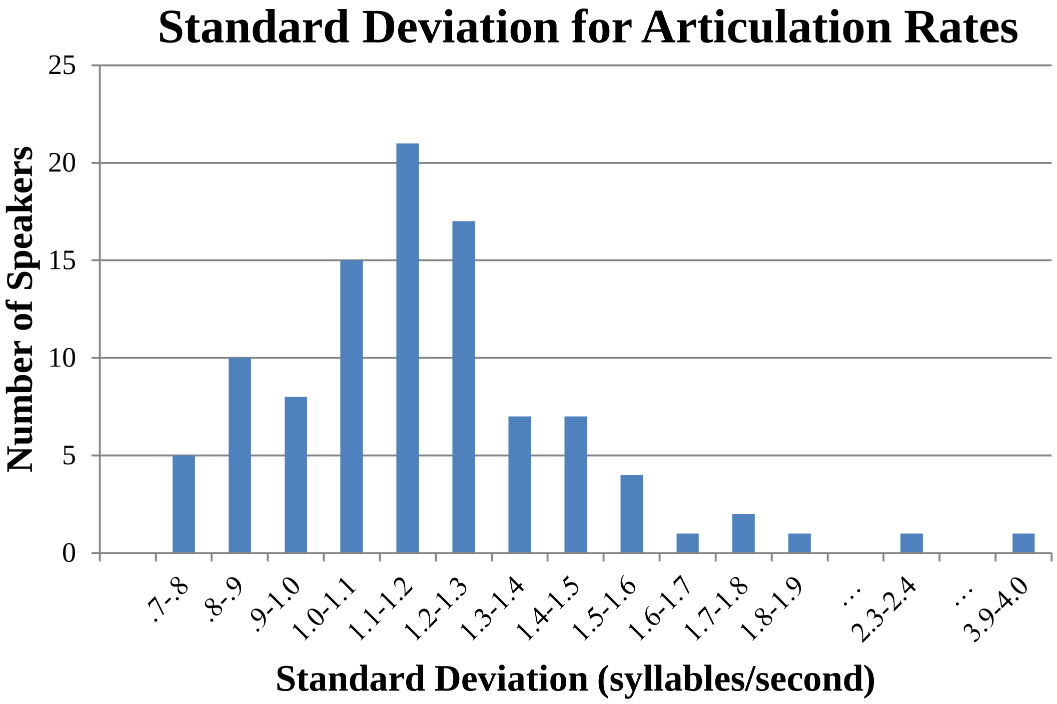


**Figure 4.2:** Distribution of standard deviations in articulation rate

The standard deviations for AR within speakers appear normally distributed. The mean SD is 1.20 syll/sec, with a range of 0.72 and 3.95 syll/sec. Those speakers who lie towards the left end of the x-axis are considered relatively more consistent in their AR than those speakers who fall towards the right end, who are characterized as having a more variable AR. The SDs of the 100 speakers lie within a range of 3.23 syll/sec, which is a larger range (by 0.01 syll/sec) than the range of means found for AR (see Figure 4.1). AR has three suspected outliers at 1.72 syll/sec, 1.77 syll/sec, and 1.87 syll/sec. There are also two extreme outliers at 2.36 syll/sec and 3.95 syll/sec.

The cumulative distribution graph of means in Figure 4.3 below shows the percentile within which a given AR falls. The y-axis is the cumulative percent of the population, and the x-axis represents AR.

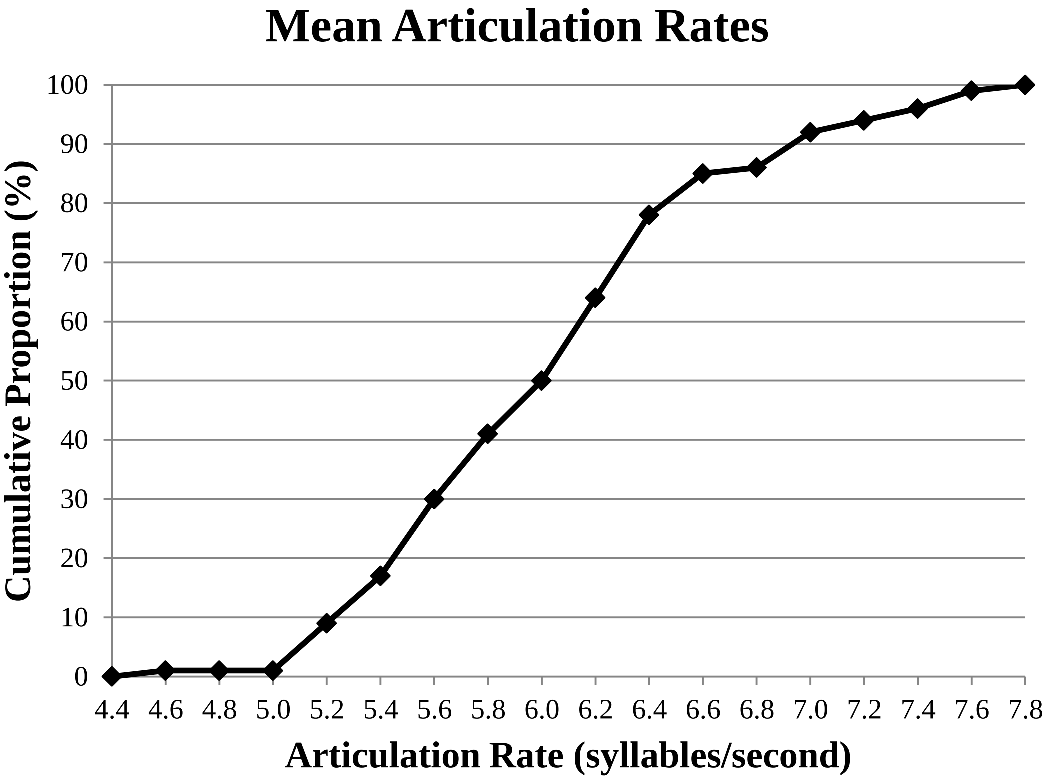


**Figure 4.3:** Cumulative percentages for mean articulation rate

The curve in Figure 4.3 is characterized by a steep central portion, but rather gentle gradients at both ends. ±1 SD from the mean gives a range of approximately 5.3 and 6.6 syll/sec, into which roughly 73% of the population falls. The cumulative distribution of individuals' SDs is illustrated in Figure 4.4, which follows the same template as that of Figure 4.3.

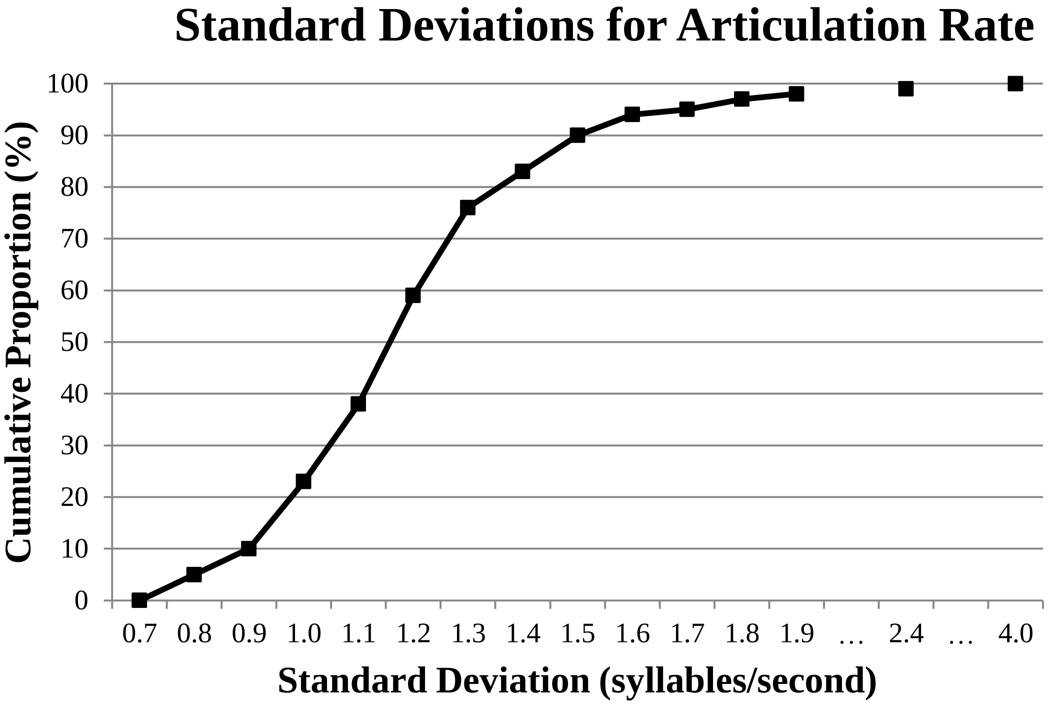


**Figure 4.4:** Cumulative percentages for standard deviation in articulation rate

The curve in Figure 4.4 is similar to the curve seen in Figure 4.3, as it is characterized by a steep central portion and gentle gradients at both ends. However, the slopes at the ends are not as gradual as the curve of the mean distributions (Figure 4.3). ±1 SD from the mean SD gives a range from approximately 0.82 to 1.58 syll/sec, a band in which roughly 84% of individuals SDs fall.

Comparing the intra-speaker variation to the inter-speaker variation, it is important to note that the mean SD (1.2 syll/sec) for a speaker is about twice the SD (0.64 syll/sec) for between-speaker variation. One would be more likely to find higher levels of variation within any given speaker of SSBE than between that speaker and others. This variability is also shown in the variance ratio, which is a calculated by dividing the squared between-speaker SD by the mean squared within-speaker SD (Rose et al., 2006). A value of less than one indicates that there is more variation within speakers than between them. A value greater than one indicates that there is more variation between speakers than within speakers. The variance ratio for AR is 0.2844, which confirms that there is more variation within individuals than there is between them.

### 4.3.4 Discussion

In addition to providing population statistics, there are three main points to be drawn from the results reported in § 4.4.0. The first is that the ARs found in the current study are very different from those found by older AR studies, namely Goldman-Eisler (1956), Robb et al. (2004), and Jacewicz et al. (2009). Goldman-Eisler (1956) reported a mean of 4.95 syllables per second, a mean range of 4.40 to 5.90, and a standard deviation of 0.91 syllables per second (range = 0.54 to 1.48). Her research was based on the spontaneous speech of eight British adults recorded in 30- to 60-minute interviews. Her method of calculating AR permitted certain disfluencies to be included in the material (e.g. unnatural sound prolongations), and this perhaps in part accounts for her lower mean AR than that found in the present study. Another reason for higher AR results in the present study could also be due to the use of phonological syllables rather than phonetic syllables. The use of phonetic syllables could potentially lead to lower ARs, as it counts only those syllables which are actually articulated by the speaker (see § 4.3.2 for the example of "did you eat yet" (4 syllables) versus "jeet yet" (2 syllables)).

The second point is that claims that forensic practitioners who took part in the survey reported in Chapter 3 made about AR being a useful speaker discriminant appear to be misguided, as AR is a weak discriminator, because there is more variation occurring within speakers than between them. For this reason, discriminating between individuals is difficult when a person has a typical mean AR. However, this is not to say that the parameter is not helpful when discriminating between speakers with lower or higher mean ARs.

The final point is that forensic phoneticians need to take care when using AR or SR in forensic speaker comparison analysis, since SR is even more variable within a speaker than between speakers (Künzel, 2007). This parameter is best used in combination with other parameters for discriminating between speakers, but may carry more weight when AR is used to discriminate individuals who fall near the outer boundaries of the distribution.

### 4.3.5 Limitations

A possible limitation of the present study is the selection of memory stretches as the speech intervals over which syllables are to be counted. Choosing a speech interval is dependent on the short-term memory of the analyst calculating the AR. This can potentially lead to high levels of variation in AR when measured by different analysts. Ideally, forensic methodologies should be robust and easily replicable across many analysts in order to achieve comparable results. For this reason it is important to verify the AR results calculated from memory stretches by comparing memory stretch interval results to those obtained from more commonly defined and objective interval (i.e. inter-pause stretches).

## 4.4 Redefining the Speech Interval

The following section compares the results found for memory stretches in § 4.3 to those found for inter-pause stretches.

### 4.4.1 Methodology

Twenty-five of the same speakers as those reported on above were randomly selected from Task 2 of DyViS, and five minutes of speech from each individual was analyzed starting at a point two minutes into each speech sample. It is important to note that for comparison purposes, the mean AR for an individual using memory stretches was only calculated using intervals from the same five-minute speech sample as was used when calculating AR using inter-pause stretches. The remaining aspects of the methodology were also kept consistent, and syllables were defined phonologically.

Inter-pause stretches are defined here as both filled and unfilled pauses that lasted 130ms or longer (Dankovičová, 1997), but were not stop closures. The interval also had to include at least five syllables. The criteria for items that were excluded from analysis in an interval were identical to those set by the exclusion rules in § 4.3.1. This meant that intervals excluded any kind of pauses, either filled or unfilled, repeated syllables, unintelligible speech, and any syllable lengthening that went beyond the phonological requirements of English. A mean of 40 intervals was measured per speaker, with a range of 26-58, and amounting to 1,011 ARs in total. The mean number of syllables per interval was 9.97, with a range of 5 to 37 syllables across all speakers.

### 4.4.2 Results

The mean ARs for both memory stretches and inter-pause stretches are presented below in Figure 4.5. The y-axis represents the AR in syll/sec and the x-axis shows the 25 randomly selected speakers from the 100 speakers in the DyViS Database.

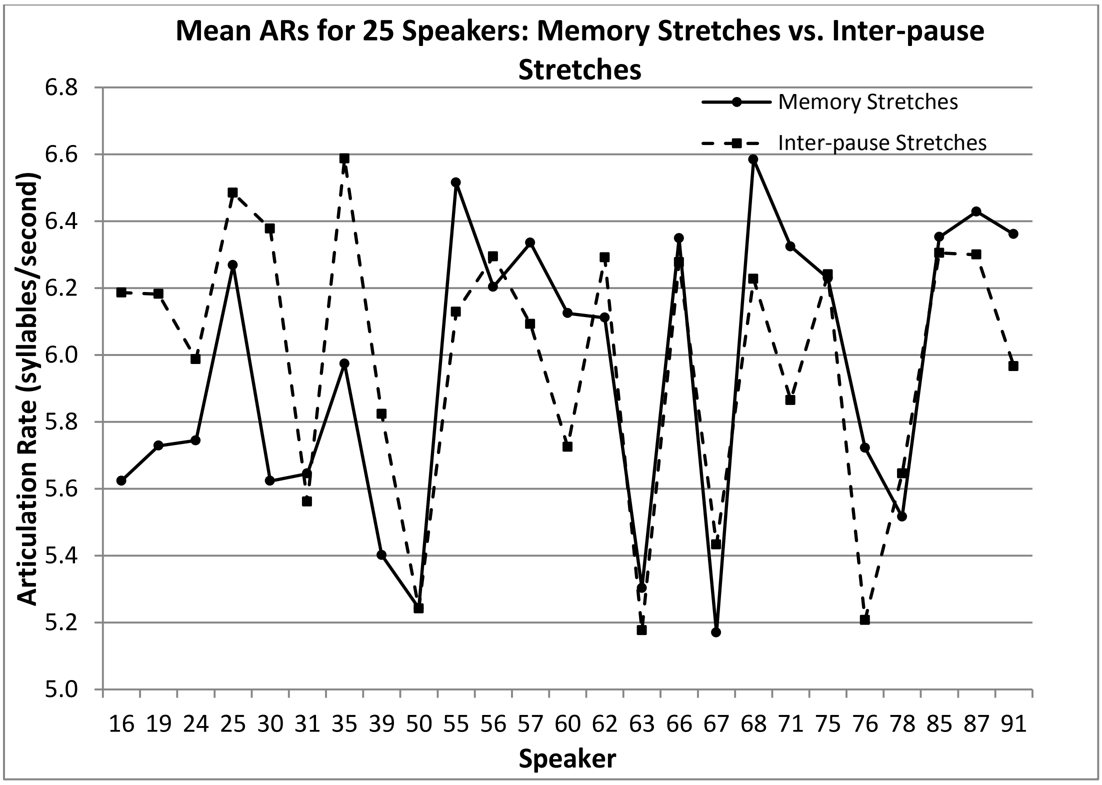


**Figure 4.5:** Comparison of mean articulation rate for memory stretches versus inter-pause stretches

Figure 4.5 provides mean ARs generated using both methodologies. The means for each speaker are displayed above one another to allow for a visual comparison of the differences found between them. The two AR methodologies prove unpredictable in terms of indicating a trend for whether one methodology produces consistently higher or lower ARs than its counterpart, as evident in the crossing lines in Figure 4.5. All speakers show relatively small differences (especially speakers 031, 050, 056, 063, 075, 078, and 085) in their mean ARs. However, some speakers have larger differences (e.g. speakers 016, 035, 068, 071, 076, and 091) than others. Using the absolute values of the differences, the average (mean) AR difference across the 25 speakers is 0.286 syll/sec, with a range of 0.001 to 0.75 syll/sec.

The mean AR for the 25 speakers using inter-pause stretches was 5.98 syll/sec, compared to the mean AR of the same speakers calculated with memory stretches, which was 5.96 syll/sec. The mean AR calculated by the two different methodologies differs by two-hundredths of a second. This is a minute amount, given that the mean ARs of the speakers are between 5.00 and 7.00 syll/sec and the mean SD for the 25 speakers (using memory stretches) was 0.64 (syll/sec). Using a Wilcoxon signed rank test for the two sets of data, the null hypothesis is retained (there is ***no*** significant difference between the two methods), as the p-value is 0.74. This provides a validation of the memory stretch method for the calculation of ARs.

### 4.4.3 Discussion

The present study gives rise to two important conclusions. The first is that AR results appear to be unaffected by the definition of the speech interval as long as the following are kept consistent: the basic unit of speech defined here as the phonological syllable, and the exclusion rules. Based on the findings in § 4.4.2 and the experience gained from calculating 125 mean ARs, I am now of the opinion that mean AR measurements are affected more by the exclusion rules than they are by the actual definition of the speech interval (memory stretch vs. inter-pause unit). The exclusion rules were described in § 4.3.2 and concern what speech can be excluded from analysis, e.g. whether false starts, unnatural prolongation, and unintelligible speech are to be included or excluded. These exclusion rules can vary from analyst to analyst, as one might find a repetition such as "I-I-I-I am going to the store" should be excluded, but something such as "I am I am going to the store" should be included. The judgments made with respect to exclusion in an analysis may be exercised by the analyst at a level below his/her conscious awareness. However, speaking from experience, an analyst remains relatively consistent working within the constraints of the rules. Therefore, the differences that were found among the mean ARs for the 25 speakers could more likely be attributed to variability that naturally occurs when taking AR measurements (eg. the precise start/stop times of intervals), and acoustic measurements in general (Harrison, 2004), rather than to a difference caused by the definition of a speech interval (i.e. memory stretches vs. inter-pause stretches).

The final conclusion to be drawn here is that reaffirmation that the results found in § 4.3 using memory stretches are valid and reliable measurements, since there was no significant difference found between the results arrived at using methodologies that incorporate memory stretches versus inter-pause stretches. Furthermore, the amount of time it takes to calculate ARs using inter-pause stretches is far greater than the time it takes to calculate ARs using memory stretches. Therefore, analysts might be advised to use memory stretches rather than inter-pause stretches in order to use time more efficiently but still calculate reliable results.

## 4.5 Manipulating the Syllable Requirements

Given that the methodology for selecting speech intervals does not appear to affect AR figures significantly, it is helpful to consider the effects that the syllable requirements of a speech interval have on AR figures. This section reports on the manipulation of the number of syllables used in a speech interval. AR was recalculated for 25 speakers using different syllable lengths for speech intervals.

### 4.5.1 Methodology

The data for the first 25 speakers of the study that were presented in § 4.3 were used to recalculate mean ARs using different minimum syllable requirements for speech intervals. There were seven possible minimum syllable requirements for the speech interval, ranging from four to ten[20]. Microsoft Excel was used to remove tokens from the speakers' data for each given requirement and mean ARs and SDs for all individuals were recalculated once tokens had been removed from each minimum syllable level.

### 4.5.2 Results

The mean AR and SD results across all speakers for different minimum syllable requirements in a speech interval are provided in Table 4.3. The table details for each minimum syllable level the mean number of tokens included for calculating speakers' overall AR, as well as the group's mean AR, and the group's mean SD.

**Table 4.3:** Summary of articulation rate statistics when varying the minimum number of syllables in the speech interval

| Syllables | Mean Number of Tokens | Mean AR (syll/sec) | Mean SD of Speakers |
|---|---|---|---|
| 4 ≤ | 29.96 | 5.738 | 1.134 |
| 5 ≤ | 29.04 | 5.762 | 1.120 |
| 6 ≤ | 26.96 | 5.844 | 1.105 |
| 7 ≤ | 24.40 | 5.929 | 1.089 |
| 8 ≤ | 20.92 | 6.024 | 1.072 |
| 9 ≤ | 17.12 | 6.096 | 1.072 |
| 10 ≤ | 13.32 | 6.129 | 0.973 |

[20] This range was chosen based on the number of tokens available for the different syllable lengths.

There is a prominent trend present in Table 4.3: as the minimum number of syllables required for a speech interval increases, the mean AR for speakers also increases while the mean SD decreases. This shows that perhaps AR becomes slightly more stable within individuals as the minimum syllable count per speech interval is increased. However, it is necessary to examine the results for changes in within-speaker variation as well as changes in between-speaker variation. Table 4.4 shows the mean differences in mean, SD, and the difference (Δ) for speakers at different minimum syllable requirement levels. The first column indicates the minimum number of syllables required for a speech interval, and the second and third columns display the mean Δ for mean AR and SD. The Δs in columns two and three are calculated by taking the average value (mean AR or SD) at a given syllable level and subtracting the baseline means (those values calculated for 4≤ syllables). Positive values represent an increase, while a negative value would indicate a decrease in AR.

**Table 4.4:** Within-speaker differences for articulation rate

| | **Within-speaker** | |
|---|---|---|
| | **Mean Δ for Mean AR (syll/sec)** | **Mean Δ for SD (syll/sec)** |
| 5 ≤ | 0.024 | -0.013 |
| 6 ≤ | 0.106 | -0.029 |
| 7 ≤ | 0.192 | -0.045 |
| 8 ≤ | 0.286 | -0.062 |
| 9 ≤ | 0.359 | -0.062 |
| 10 ≤ | 0.392 | -0.160 |

Table 4.4 shows that individuals are patterning similarly to the group as a whole in that mean AR increases and SD decreases as the minimum number of syllables in a speech interval is increased. It appears that the higher the

minimum number of syllables in a speech interval the more stable a speaker's AR becomes. This could also be due in part to the decreasing number of tokens involved in the calculation at the 10≤ syllable level. However, the number of tokens is still rather robust for syllables at the nine or more level and below, across which there is still a demonstrable decrease in within-speaker variation. Table 4.5 examines between-speaker differences found for different minimum syllable requirements. The first column indicates the minimum number of syllables required for a speech interval, and the second presents the mean Δ for the SD of the mean ARs. As in Table 4.4, a positive value shows an increase, while a negative value shows a decrease in AR.

**Table 4.5:** Between-speaker differences for articulation rate

| | **Between-speaker** |
|---|---|
| | **Mean Δ for SD of AR Means (syllables/second)** |
| 5 ≤ | -0.022 |
| 6 ≤ | -0.004 |
| 7 ≤ | -0.002 |
| 8 ≤ | 0.014 |
| 9 ≤ | -0.006 |
| 10 ≤ | 0.028 |

The overall trend displayed in Figure 4.8 is inconsistent, in that as the minimum syllable requirement is increased the SD of the AR means fluctuates (both increases and decreases to different degrees). Across all minimum syllable requirement levels the mean Δ is 0.001 syll/sec. This means that the variation between speakers' mean ARs remains relatively stable despite the changes made to the minimum number of syllables required in a speech interval.

### 4.5.3 Discussion

The conclusion to be drawn from the results obtained by increasing the minimum number of syllables required in a speech interval is that within-speaker variation will decrease while the between-speaker variation remains rather unchanged. This is important for the use of AR in an approach utilizing likelihood ratios and could potentially lead to stronger evidential values for AR as a discriminant. In the next section, LRs are calculated for AR, as well as to ascertain whether increasing the minimum number of syllables in a speech interval improves the strength of evidence for AR.

## 4.6 Likelihood Ratios

LRs provide a framework for the estimation of the strength of evidence under competing defense and prosecution hypotheses (see Chapter 2). In order to assess the discriminatory power of AR numerically, same speaker (SS) and different (DS) LRs are calculated and plotted.

### 4.6.1 Methodology

The LR calculations for AR were performed using a MatLab implementation of Aitken and Lucy's (2004) Multivariate Kernel-Density (MVKD) formula (Morrison, 2007). MVKD was chosen over Lindley's (1977) univariate LR, because it can account for inter- and intra-speaker variation (i.e. using kernel-densities), whereas Lindley's (1977) formula makes LR calculations based on normally-distributed data (i.e. a single distribution curve) for both the LR numerator and denominator. Morrison also suggests that because Lindley's LR formula cannot account for "occasion-dependent within-speaker variation" (Morrison, 2008, p. 97), the MVKD approach is more suitable for analyzing speech evidence, although it is not ideal.

The MVKD formula provided by Aitken and Lucy (2004) assumes that within-speaker variability is normally distributed (numerator). The between-speaker variation, however, is not assumed to be distributed normally and is estimated using kernel-density, a measure which accounts for skewed distributions. The Gaussian mixture model - universal background model (GMM-UBM[21]; Reynolds et al., 2000) has also been proposed for calculating LRs in ASR, and has been used to calculate LRs in phonetic/linguistic-based FSCs (Becker et al., 2008; French et al., 2012; Rose and Winter, 2010). GMM-UBM, like MVKD, can accommodate multivariate data; however, GMM-UBM models the data differently from MVKD. GMM-UBM utilizes GMMs to characterize distributions instead of kernel densities (as per MVKD). The most significant difference between GMM-UBM and MVKD is that the MAP (maximum a posteriori; Reynolds et al., 2000) background model for GMM-UBM is person-independent and is compared against a model of person-specific parameter characteristics when comparing same (SS) and different (DS) speaker pairs (Reynolds et al., 2000). Morrison has found that a GMM-UBM, which does not assume normal distribution for within- or between-speaker values, performed both better and worse than MVKD on different occasions (Lindh and Morrison, 2011; Morrison, 2011). However, the application of the GMM-UBM has predominantly been tested using automatic systems. Despite these findings,

[21] GMM refers to the way in which the data are modeled. GMM is a parametric density function that is comprised of a number of component functions (Gaussian; Reynolds et al., 2000). UBM refers to the way in which the background population in modeled. A UBM is used to represent general, person-independent parameters to be compared against a model of person-specific parameters (Reynolds et al., 2000).

MVKD has been shown to provide reliable and important strength-of-evidence results across many studies using more traditional acoustic-phonetic features (Hughes, 2011; Morrison, 2008; 2009a; Rose, 2006a; 2007a, 2007b), and will therefore be applied here to AR.

A MatLab script (ss_ds_lrs.m)[22] was used to run multiple same-speaker (SS) and different-speaker (DS) LR calculations for AR. The script calls for the 100 speakers' samples to be split in half (i.e. 50/50), such that SS comparisons may be performed (50 SS comparisons), which in turn results in 2,450 DS comparisons (50*49). Speakers 001-050 act as the speaker comparisons, while speakers 051-100 act as the background population. The calculated raw LRs were transformed using natural and base$_{10}$ logarithms. The transformation, allows zero, rather than one, to act as the center point between the support for $H_p$ and $H_d$. The log transforms are also beneficial in normalizing distributions which may be skewed by large and infrequent values.

The magnitudes of LRs are discussed and assessed with reference to Table 4.6. The verbal scale, adapted from Champod and Evett (2000), is based on identifying a Log$_{10}$ LR that corresponds to a verbal expression representing the strength of evidence in favor of the prosecution hypothesis ($H_p$: same speaker) or the defense hypothesis ($H_d$: different speakers).

[22] The script was developed by Phil Harrison from J P French Associates.

**Table 4.6:** Expressions for strength of evidence in terms of $\text{Log}_{10}$ LR and the corresponding verbal expression following Champod and Evett's (2000) verbal scale

| $\text{Log}_{10}$ LR | Verbal expression |
|---|---|
| <-4 | Very strong evidence to support $H_d$ |
| -4 to -3 | Strong evidence to support $H_d$ |
| -3 to -2 | Moderately strong evidence to support $H_d$ |
| -2 to -1 | Moderate evidence to support $H_d$ |
| -1 to 0 | Limited evidence to support $H_d$ |
| 0 to 1 | Limited evidence to support $H_p$ |
| 1 to 2 | Moderate evidence to support $H_p$ |
| 2 to 3 | Moderately strong evidence to support $H_p$ |
| 3 to 4 | Strong evidence to support $H_p$ |
| > 4 | Very strong evidence to support $H_p$ |

This scale was previously used by the UK Forensic Science Service (Champod and Evett, 2000), and will serve as a means of strength-of-evidence evaluation for the remainder of the thesis.

Performance of the system is discussed in respect of log-LR cost (Cllr) and equal error rate (EER), which are both metrics of system validity. The term "validity" refers to how well a system (in this thesis a system can be an individual parameter or multiple parameters combined) can distinguish between same-speaker (SS) and different-speaker (DS) pairs. Severity of performance error was assessed using Cllr, which is a common assessment used in automatic speaker recognition/comparison (Ramos Castro, 2007). The Cllr is a Bayesian error metric that quantifies the ability of the system to output LRs that align correctly with the prior knowledge of whether speech samples were produced by the same or different speakers. The Cllr acts as an error measure that captures the "gradient goodness of a set of likelihood ratios derived from test data" (Morrison, 2009b, p. 6). Previous studies of LRs for forensic speaker comparisons have shown that Cllr proves appropriate for measuring errors

(Morrison and Kinoshita, 2008; Morrison, 2011). The equation commonly used for calculating Cllr is provided in Equation (1).

(1)

$$C_{\mathrm{llr}} = \frac{1}{2}\left( \frac{1}{N_{\mathrm{ss}}} \sum_{i=1}^{N_{\mathrm{ss}}} \log_2\left(1 + \frac{1}{\mathrm{LR}_{\mathrm{ss}_i}}\right) + \frac{1}{N_{\mathrm{ds}}} \sum_{j=1}^{N_{\mathrm{ds}}} \log_2(1 + \mathrm{LR}_{\mathrm{ds}_j}) \right)$$

**from Morrison (2009, p. 2391)**

$N_{ss}$ = Number of same speaker pairs
$N_{ds}$ = Number of different speaker pairs
$LR_{ss}$ = LR from same speaker pairs
$LR_{ds}$ = LR from different speaker pairs

Cllr was calculated using Brümmer's FOCAL toolkit[23] function *cllr.m* with the log-LRs as input. Values of Cllr that are closer to zero indicate that error is low. For values approaching one the error is considered to indicate poor performance, while values above one indicate very poor performance (van Leeuwen and Brümmer, 2007, pp. 343-344).

EER, unlike Cllr, provides a "hard" (i.e. binary) accept-reject measure of validity. This is based on the point at which the percentage of false hits (DS pairs that ostensibly offer support for the $H_p$) and the percentage of misses (SS pairs that appear to offer support for the $H_d$) are equal (Brümmer and du Preez, 2006, p. 230).

### 4.6.2 Results

The following sections detail the results of two sets of calculations of the discriminant potential of AR. The first investigates the capacity of AR to discriminate within a large homogeneous group of 100 males. The second

[23] http://sites.google.com/site/nikobrummer/focal (Downloaded: 13 August 2012)

considers how the minimum number of syllables in a speech interval may affect the strength-of-evidence for AR.

#### 4.6.2.1 LRs for ARs in 100 SSBE Male Speakers

The results for the calculation of LRs on ARs are summarized in Table 4.7. The second row contains the results from SS comparisons and the third row contains DS comparison results. The total percent of correct SS and DS comparisons is shown in the second column. Correct LRs are determined by whether or not an LLR for SS comparisons is a positive value (providing support for the prosecution hypothesis) and whether an LLR for DS comparisons is a negative value (providing support for the defense hypothesis). The third through fifth columns report on the strength-of-evidence, whereby the third column presents the mean LLR for all comparisons (either SS or DS). The smallest calculated LLR is in the fourth column, followed by the largest calculated LLR. The final two columns provide the EER and Cllr values for the entire system.

**Table 4.7:** Summary of LR-based discrimination for articulation rate (100 speakers)

| Comparisons | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **AR SS** | 90.0 | 0.18 | -1.48 | 2.06 | .3340 | .8981 |
| **AR DS** | 46.2 | -2.94 | -8.76 | 0.82 | | |

Table 4.7 shows that AR performs much better with SS comparisons than DS comparisons. The results may seem counterintuitive, since there is higher within-speaker variability than between-speaker variability for AR, and it might be assumed that the high within-speaker variation would cause DS pairs to perform better than SS pairs. However, it appears that, because the degree of

variation in AR is so high within speakers overall, the system tends to allocate higher degrees of similarity if two speakers have similar degrees of (high) within-speaker variation. This is evident in the fact that for DS comparisons, the system performs slightly worse than chance (which is 50%, since an LLR correct/incorrect response is categorized as supporting the $H_p$ or the $H_d$, respectively) as the AR system tends to over-predict pairs being from the same speaker rather than different speakers (note the high error rate in correct DS judgments) [24]. Table 4.7 shows that Cllr is approaching one, but is still under it. Following van Leeuwen and Brümmer (2007) this would be classified as a 'poor' performance. The EER is high at 33.4% for AR as a system, and the mean SS LR offers only limited evidence to support the prosecution hypothesis ($H_p$). The mean DS LR is slightly stronger, and offers moderate evidence to support the defense hypothesis ($H_d$).

The Tippett plot in Figure 4.6 provides a visual measure of the performance of AR as a discriminant feature. The *x-axis* displays $\log_{10}$ LRs where zero is the division between support for $H_p$ (>0) and support for $H_d$ (<0). The y-axis displays cumulative proportion. Flatter contours indicate a higher proportion of pairs that achieve a stronger strength-of-evidence, and contours that are steeper indicate a weaker strength-of-evidence. The results for SS and DS comparisons are assessed together.

[24] An additional explanation for the poor performance of DS comparisons could be that the system is not optimally calibrated (see § 8.5.3). This is evident in the intersection between the SS and the DS distributions in Figure 4.6, as the intersection is not at LLR = 0, but further to the right into the higher scores. Therefore, many DS comparisons obtain an LLR larger than zero. This miscalibration is also potentially the reason for the poor Cllr values.

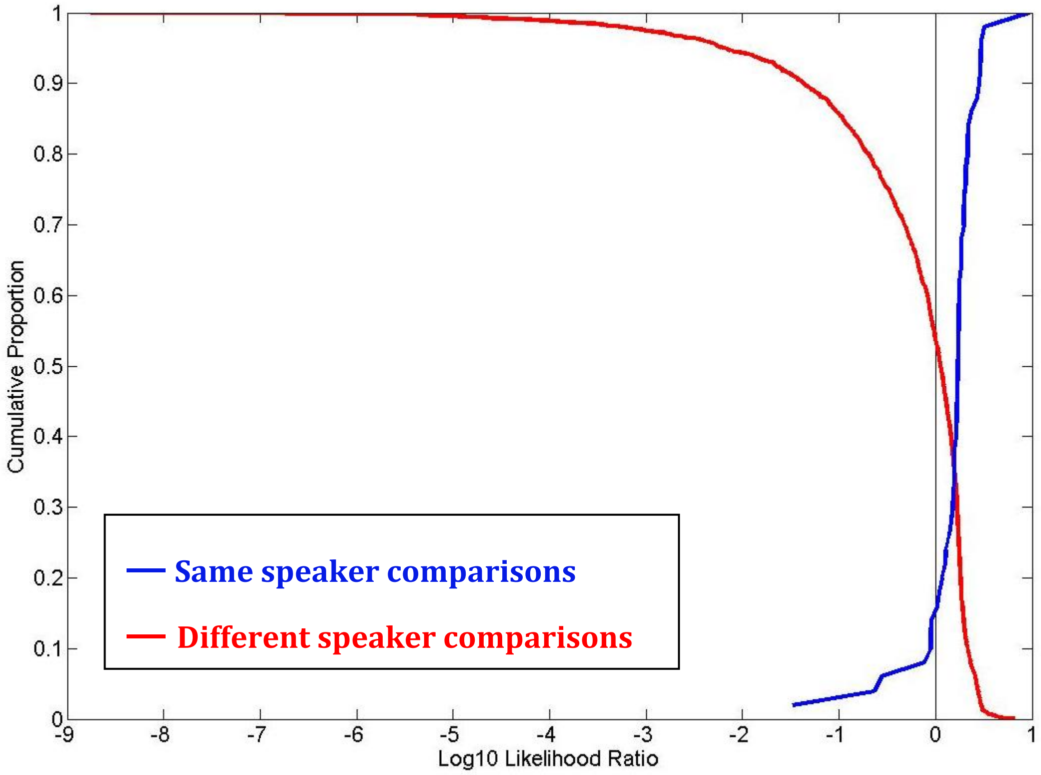


**Figure 4.6:** Tippett plot of articulation rate

Figure 4.6 shows that error rates are higher for DS comparisons than they are for SS comparisons. The SS line is steeper than that of the DS line and provides a relatively low strength of evidence. DS, on the other hand, can attain higher strength of evidence (a $\text{Log}_{10}$ LR of -5 or even lower), although these values are reserved for a very small percentage of DS comparisons. It is important to remember when analyzing SS and DS LR results that “two samples cannot get more similar for a feature than identical” (Rose et al., 2006, p. 334), and therefore DS comparisons will always carry the potential for achieving a higher strength of evidence than SS comparisons. The Tippett plot paints an overall picture that AR as an individual parameter is relatively weak at discriminating between individuals, and only produces higher strength of evidence for a very small proportion of DS comparisons.

### 4.6.2.2 LRs for ARs of 25 Speakers with Variation in the Minimum Number of Syllables in a Speech Interval

The following section reports on the LRs calculated for the ARs of 25 speakers while varying the minimum number of syllables required in a speech interval. Table 4.8, which has a similar structure to that in Table 4.7, provides the results of the different systems. The first column includes a value next to the SS or DS lines to indicate the minimum number of syllables required for given speech intervals in that system.

**Table 4.8:** Summary of LR-based discrimination for mean articulation rate when varying the minimum number of syllables in a speech interval (25 speakers)

| Comparisons | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **4 ≤ Same Speaker** | 84.6 | 0.005 | -1.800 | 0.369 | 0.2500 | 1.0260 |
| **4 ≤ Different Speaker** | 52.6 | -0.252 | -3.153 | 0.771 | | |
| **5 ≤ Same Speaker** | 76.9 | -0.014 | -1.703 | 0.340 | 0.3109 | 1.0326 |
| **5 ≤ Different Speaker** | 53.2 | -0.260 | -2.676 | 0.665 | | |
| **6 ≤ Same Speaker** | 69.2 | -0.068 | -1.73 | 0.412 | 0.3846 | 1.0976 |
| **6 ≤ Different Speaker** | 54.5 | -0.316 | -2.718 | 0.608 | | |
| **7 ≤ Same Speaker** | 53.8 | -0.152 | -1.908 | 0.435 | 0.4615 | 1.1780 |
| **7 ≤ Different Speaker** | 55.1 | -0.345 | -2.751 | 0.585 | | |
| **8 ≤ Same Speaker** | 69.2 | -0.103 | -2.731 | 0.662 | 0.3782 | 1.1104 |
| **8 ≤ Different Speaker** | 53.8 | -0.280 | -1.839 | 0.409 | | |
| **9 ≤ Same Speaker** | 61.5 | -0.135 | -1.243 | 0.414 | 0.4615 | 0.9958 |
| **9 ≤ Different Speaker** | 52.6 | -0.025 | -0.595 | 0.317 | | |
| **10 ≤ Same Speaker** | 53.8 | 0.024 | -1.175 | 0.322 | 0.5385 | 0.9848 |
| **10 ≤ Different Speaker** | 39.7 | -0.033 | -0.234 | 0.218 | | |

Increasing the minimum number of syllables for a speech interval in the different systems did not improve the percentage of correct SS or DS comparisons; however, it does improve the Cllrs, as seen in the rightmost column in Table 4.8. By increasing the minimum number of syllables required for a speech interval, the percentage of correct SS comparisons drops by 30.8%. For DS comparisons that number drops by 15.4%, with Cllr improving by .0412.

EER is more erratic, showing both increases and decreases in its performance as the minimum syllable number in an interval is increased (although it is mainly an upward trend).

In comparison to the system discussed in § 4.6.2.1 for 100 speakers, the percentage of correct SS comparisons and the Cllr value are worse, while the percentage of correct DS comparisons is slightly improved. Overall, EERs for minimum syllable lengths above five perform worse than the AR system in § 4.6.2.1. Caution should be exercised when interpreting these results, however, as there were only 13 test speakers and 12 reference speakers, whereas the tests described in § 4.6.2.1 included 50 speakers in both test and reference sets.

### 4.6.3 Discussion

Although within-speaker variation can be decreased by increasing the minimum number of syllables in a speech sample, it appears that the system performs worse overall in terms of the percentage of speaker comparisons judged correctly, in spite of the Cllr improving slightly. However, no system produced a Cllr better than that seen in § 4.6.2 with all 100 speakers and a minimum of four syllables per speech interval. It is important to note that, by increasing the minimum number of syllables required in a speech interval for the 25 speakers, the number of useable speech intervals available for each individual is decreased, potentially affecting the results. This means that systems were potentially performing worse due to a decrease in the number of speech intervals available; or rather, the change is the result of a combination of this along with the increase in the minimum syllables required for a speech interval.

Most importantly, by re-redefining the minimum syllable count requirement for a speech interval the overall performance of a system is susceptible to changes in data collection and methodologies for AR. The need for consistency in analysis techniques for all forensic domains, including AR in forensic speech science, may be best achieved through prescribed methods (based on rigorous empirical testing) for calculation.

## 4.7 Conclusion

Overall, it appears that AR can be classified as a speech parameter that carries higher intra-speaker variation than it does inter-speaker variation. In respect of the large number of available methodologies for the calculation of AR, it appears that the defining of a speech interval (memory stretch vs. inter-pause unit) does not have a significant effect on the results. However, in the context of real forensic casework if methodologies and analysts are to be kept consistent for the analysis of suspect and criminal samples then problems relating to AR calculation methods will be minimized.

AR as a discriminant parameter has proved to be a very poor one[25], and it is not anywhere close to being as good at discriminating between individuals as experts have claimed it to be (Chapter 3). This raises the question as to why some analysts are using it at all in casework except for instances of very high or low AR. However, exceptions exist for those speakers that are classified as outliers. It has been shown that AR offers a very weak strength of evidence for SS comparisons, while DS comparisons can potentially offer a higher level of strength of evidence. However, this must be traded off against the fact that they

[25] At least as these results suggest.

produce a very high rate of incorrect DS judgments. Despite all efforts to decrease within-speaker variation by increasing the minimum number of syllables in a speech interval, the overall system is not improved from the original system results shown in § 4.6.2. The Cllr in the best-performing system is .8981, which classifies it as performing poorly since the score is close to 1, following Brümmer and du Preez (2006).

The analysis of AR as a parameter under an LR framework in forensic speech science urges caution for casework, in that parameters previously thought to be good speaker discriminants might transpire to carry higher intra-speaker variation than inter-speaker variation, which will generally result in a lower strength of evidence for a given parameter. More research on speaker discriminants for other commonly-used parameters in forensic casework is clearly needed, because there is a risk that some experts in the field are analyzing certain features rather blindly. That is to say, they are giving weight to features which actually provide little in terms of discrimination power. This is shown by the fact that 93% of experts surveyed analyze speech tempo, and 73% of those do with varying regularity; furthermore, 20% of experts reported speech tempo to be the single most useful discriminant. The analysis carried out in this chapter provides evidence that AR may be a far from useful discriminator in many cases.

Although AR may not be the discriminant shibboleth all experts hope for, it is important that AR is still considered in forensic speaker comparisons in conjunction with other speech parameters. There are instances in which speakers may have a very low or high AR, and in which the parameter can be considered useful (either as evidence for or against speaker identity). As Rose

(2007a, p. 1820) points out, “not all speakers differ from each other in the same way”.  Therefore, there will be a few individuals for whom AR is potentially a good discriminant parameter, as was evident in the small number of high LRs for DS comparisons in § 4.6.

# Chapter 5: Long-Term Formant Distributions

## 5.1 Introduction

Forensic speech science literature is largely characterized by research investigating the discriminant power of vowels. In Chapter 3, it was reported that 28% of forensic phoneticians found vowels to be the most useful feature in discriminating between speakers; overall, this places vowels as the second highest-ranked discriminant parameter (along with accent/dialect variants) among all possible parameters analyzed for FSCs. Long-term formant (frequency) distribution (LTFD) work to date constitutes only a small portion of the research carried out. LTFD is the method used to calculate the average values for each formant of a speaker over a given speech recording. For a given formant (i.e. F1-F4), measurements for all vowels produced by a single speaker are averaged across the entire recording or relevant portions of the recording. This means that for each formant (F1-F4) of a speaker there is an LTFD value and a standard deviation (SD), which will be referred to as LTFD1, LTFD2, LTFD3, and LTFD4. LTFDs are frame-by-frame measurements (5 msecs in length for the current study); therefore, long vowels carry more weight than short vowels in that they yield a greater number of measurements per vowel. A positive attribute of LTFDs is that they do not require the categorization of individual vowels into phoneme classes, as all vowels are considered in an analysis. This results in greater time savings. LTFD also avoids the potential correlations between vowel phonemes which would not allow those correlated phonemes to be combined, since combinations of parameters under Bayes' theorem are only allowed if events can be shown to be mutually exclusive.

This chapter provides population statistics for LTFD while also investigating the discriminant power of LTFD under a likelihood ratio framework. Population statistics which consist of LTFD means and SDs are reported for F1-F4, in order to give indications of between- and within-speaker variation. LRs are calculated for LTFD1-4 individually as well as in different combinations relevant for casework. The results of the LRs are presented and considered in terms of strength of evidence, Cllr, EER, and proportion of SS and DS comparisons that were correctly identified.

## 5.2 Literature Review

The results of vowel research in forensic phonetics are well documented in the literature as a way in which to characterize the speech of an individual. Many different methods have been offered for acoustic analysis, the most common being temporal mid-point center-frequency measurements of formants for different vowels (Jessen, 2008; Rose, 2002; 2006a, b, c; Rose et al., 2003). Investigations have also been carried out using formant dynamics in order to capture the trajectories of specific vowels (McDougall, 2004; McDougall and Nolan, 2007; Hughes, 2011; Hughes et al., 2009). Formant dynamic research revealed that formants appear to be consistent within speakers and that there is variation with respect to formants between speakers. This research led to the argument that the development of techniques for measuring dynamic features should be given more attention (McDougall, 2006). Long-term spectra (LTS) were developed with this in mind, as a means of capturing the average of all spectral slices in a recording. However, LTS considers voiced speech as well as voiceless portions which could potentially include background noise (Nolan and Grigoras, 2005). LTFD, like LTS, was also developed to identify the dynamism of formants, but it is only concerned with the vowels and certain voiced portions in a recording (Nolan and Grigoras, 2005). There have also been a number of analyses concerned with vowels under a LR framework (e.g. Alderman, 2004; Kinoshita and Osanai, 2006; Rose, 2007a; Morrison, 2009). However, only three have considered LTFD (Becker et al., 2008; French et al., 2012; Jessen et al., 2013).

Nolan and Grigoras (2005) were the first to report the use of LTFD for forensic speaker comparisons. In their study, the authors consider analysis of LTFD1 and LTFD2 in order to eliminate a suspect who is thought to have made some obscene phone calls. The first author carried out an auditory analysis of vowels and took mid-point center-frequency measurements of monophthongs. Diphthongs were also included in the analysis and the beginning and end points of the vowels were measured. In general, the vowel analysis by the first author suggested that the speech in the criminal samples and the speech in the suspect sample were poorly matched. Each vowel in the suspect samples exhibited systematic differences from those found in the criminal samples, thereby rendering the criminal and suspect samples incompatible. Given that there were other equally valid methods to arrive at acoustic characterization of speakers, the second author carried out a re-analysis using alternative approaches.

The second author had previously developed new techniques for speaker comparison (Grigoras, 2001; 2003) that included speaking fundamental frequency (SFF), LTS, and LTFD. All three approaches were carried out using Catalina Toolbox[26]. LTFD was used in the re-analysis of the data in the obscene phone call case. This approach considers the "long-term disposition of formants" (Nolan and Grigoras, 2005, p. 162), an aspect which LTS fails to grasp. Only the voiced frames in the recordings were used for analysis and linear prediction was used to estimate LTFD1-4. The results from LTFD1-4 give an overall indication of each formant distribution, from which it is clear that LTFD2 and LTFD3 in the criminal recordings are considerably lower than in the suspect recording. LTFD4 is also shown to be relatively higher in the criminal samples than the suspect sample. Given the distribution of LTFD, this analysis gave further substance to the argument that there were two different speakers involved and that the suspect and criminal were not one and the same person. With respect to the approaches employed in the re-analysis, LTFD provided a "very clear picture of the average behaviour of each formant" (Nolan and Grigoras, 2005, p. 169). It also provides strong insights into the dimensions of a speakers' vocal tract, where these are reflected in the maximum LTFD[27]. The formant frequency values for LTFD are inversely related to the speaker's vocal tract size, whereby a longer vocal tract will result in lower formant values (Nolan and Grigoras, 2005; French et al., 2012). LTFD also has the capacity to indicate certain habits speakers use, such as palatalization, which are indicated by a raised LTFD2 (Nolan and Grigoras, 2005). French et al. (2012) also show that voice qualities related to tongue body position are correlated with LTFD. Additionally, the shape of the distributions for the estimates of each formant is useful in identifying speakers who have either more or less variable formants.

[26] Available at http://www.forensicav.ro/download.htm

[27] The maximum LTFD (for LTFD1 and LTFD2) is reflected in the overall area of a speaker's vowel space.

This is classified by distributions which are leptokurtic (narrow-peaked) or platykurtic (broad-peaked). Although LTFD provides promising characterizations of individuals' speech, it fails to reveal the variation that exists in specific vowel segments.

Moos (2010) utilized the LTFD method detailed in Nolan and Grigoras (2005) and analyzed LTFD2 and LTFD3 values of mobile phone speech in both read and spontaneous speech of 71 male German speakers from the Pool 2010 corpus (Jessen et al., 2005). The spontaneous speech was elicited from speakers while having them describe objects to another person (their "compatriot") without using certain proscribed words, similar to the strategies in the board game "Taboo". The person who was matched with the speaker feigned ignorance in the exchange in order to encourage more thorough descriptions and longer stretches of speech. The read speech was produced by speakers reading a German version of "The North Wind and the Sun" (Moos, 2010). Recordings were edited to include only the vocalic portions, where laterals, approximants, vocalic hesitations, and creaky voice were part of that stream. Nasals, areas of strong nasality, and vowels spoken on a high pitch (where individual harmonics were visible in the spectrogram rather than formants) were not included. After cutting down the recordings in Wavesurfer, the length of the spontaneous speech was between 12 and 83 seconds (mean = 40 sec) and the length of the read speech was between 8 and 16 seconds (mean = 12 sec).

Moos (2010) found LTFD3 values to be slightly more helpful[28] than LTFD2 in terms of speaker characterization because, overall, the former had

[28] This is also seen in Simpson (2008) and Clermont et al. (2008) with regard to F3 values for a number of phonemes that were measured using mid-point center frequencies.

smaller intra-speaker variation. The LTFD values from read speech were reported as being higher than those in spontaneous speech. However, this could be due in part to the fact that the "The North Wind and the Sun" is a tool used in phonetic experiments to get an array of phonemes and tokens. For this reason, the vowel spaces of speakers could be artificially enlarged, returning a greater spread of LTFD values per speaker. It is important to note that details of the spontaneous speech that was elicited were not provided, and could have in theory given a similar distribution of phonemes and tokens. However, that is unlikely, as spontaneous speech tends not to provide as wide of an array of phonemes as is usually the case with read speech (of course this is dependent on the chosen text). It was also noted by Moos (2010) that it is vital to know whether there is a sufficient amount of data in order to analyze LTFD; 6 seconds of pure vocalic stream were suggested as a minimum. Overall, Moos (2010) classifies LTFD as a valuable measure to include in forensic speaker comparisons and identifications. LTFD was also found to be independent of (i.e. not correlated to) F0, dialect, and speech rate, making LTFD viable for combination with these parameters under an LR conclusion framework.

Becker et al. (2008) investigated the use of Gaussian Mixture Models for LTFD1-3 under an LR framework (see Jessen et al., 2013 for a similar LTFD study but using the software Vocalise[29]). Spontaneous speech was used from 68 male German speakers from the Pool 2010 corpus recorded in a laboratory setting. The speech had been elicited as described above in Moos (2010), and as in Moos (2010), the data used in Becker et al. (2008) were transmitted through mobile phone connections in order to simulate forensically-relevant recordings.

[29] http://www.oxfordwave research.com/j2/products/vocalize [ Accessed: 8 August 2013]

These recordings were then edited to remove consonantal information as well as portions of speech where formant structures were not clearly visible. Formant tracking was then used to identify peaks and LTFD measurements were extracted in Wavesurfer. The formant measurements for the first half of each recording were used as a training set, and the test set consisted of the second half of the formant measurements. Those formant measurements in the test set were halved again in order to increase the possible number of comparisons. This resulted in recordings from the training set being around 22 seconds in length, while those in the test set were around 11 seconds long. LTFD1-3 as well as their corresponding bandwidths were considered in the analysis. 18 speakers' measurements were used to create the Universal Background Model (UBM), and one Gaussian Mixture Model (GMM) was estimated for the reference population, using 8 mixtures. The remaining 50 speakers were used in the test and a total of 100 same-speaker comparisons and 4,900 different speaker comparisons were carried out.

The lowest (i.e. the best performing) EERs were found for combinations that included bandwidths (BW), these being LTFD1+LTFD2+LTFD3+BW1+BW2+BW3 and LTFD1+LTFD2+BW1+BW2, which achieved EERs of 0.030 and 0.042, respectively. The lowest EER in which BW was not included was the combination of LTFD1+LTFD2+LTFD3, which had an EER of 0.053. Overall, discrimination levels were high, and Becker et al. (2008) note that the speaker models created using LTFD can relate directly to the configuration of the vocal tract, in turn perhaps revealing speaker-specific variations in the distributions.

Jessen and Becker (2010) build on Becker et al. (2008) and investigate LTFD as a speaker discriminant further. They first consider the relationship between LTFD2-3 and body height for 81 male speakers from the Pool 2010 corpus of telephone transmitted speech. Both LTFD2 and LTFD3 were found to have significant negative correlations with stature, where taller individuals were associated with lower LTFD2 and LTFD3. The Pearson correlation coefficients were both just over 30% (r = -0.316 and -0.339 respectively), but were nonetheless significant at the 1% level. Jessen and Becker (2010) then examined the consistency with which analysts measure LTFD. Five phoneticians measured LTFD2 and LTFD3 for 20 speakers from the Digs dialect corpus (Jessen and Becker, 2010). LTFD means were compared across analysts and it was found that LTFD2 had Pearson correlations between 0.84 and 0.95, while for LTFD3 these figures were between 0.98 and 0.99[30]. Consistency across analysts was higher for LTFD3 than for LTFD2, but both formants achieved highly consistent results overall, showing that the methodology for LTFD analysis is easily replicable. It has been posited that LTFD is potentially language-independent, as all vowel phonemes are averaged (Nolan and Grigoras, 2005). Jessen and Becker (2010) tested this hypothesis using three speakers of different German dialects in the Digs dialect corpus, as well as Russian and Albanian speakers under analogous recording conditions. They found that the different languages did not appear to differ in terms of the LTFD-space that they occupy (one-way ANOVA [F(4,55) = 0.44; p = 0.77]).

[30] It is unclear if these figures refer to *r* or $r^2$ values.

The authors then investigated the effects of Lombard speech[31] on LTFD. Mean LTFDs from 31 speakers in the Pool 2010 corpus (telephone-transmitted speech) were used to analyze possible Lombard effects on LTFD1-3. LTFD1 was found to be consistently higher in the Lombard condition than in the modal condition, with high levels of intra-speaker variation present. LTFD2 and LTFD3 were inconsistent in their effects across speakers, and both yielded non-significant differences. Although LTFD1 was shown to be affected by Lombard speech, it is often of limited use in forensic casework, due to the effect of telephone transmission on F1 (Byrne and Foulkes, 2004; Jessen and Becker, 2010; Künzel, 2001). Finally, the authors tested the performance of LTFD analysis modeled using GMMs against ASR. They found ASR to outperform LTFD analysis, with an EER of 0.107 for ASR as compared to an EER of 0.243 for LTFD. Logistic-regression fusion (which accounts for correlations between resulting LRs and then applies statistical weightings) was also used to try to improve results (EER= 0.108). However, it still performed worse than ASR on its own. The results found by Jessen and Becker (2010) were promising for LTFD use in forensic casework. LTFD revealed a negative correlation of individual formants with body height, a high consistency in measurements across different phoneticians, the potential to use LTFD statistics from one language across many languages, and the limited effect of Lombard speech on LTFD2 and LTFD3.

Most recently, French et al. (2012) examined LTFD in conjunction with Mel-Frequency Cepstral Coefficients (MFCCs; abstract properties of the acoustic

[31] Lombard speech is the tendency of speakers to increase their vocal effort when speaking (typically due to loud noise; Lombard (1911)).

signal, which are reflections of the dimensions of the vocal tract) and voice quality (VQ). They considered the efficacy and limitations of the three parameters, and correlations among the parameters. The study used the recordings from Task 2 of the DyViS database (Nolan et al., 2009). All original recordings were 15 to 25 minutes in duration. The recordings were edited to a minimum of 50 seconds of vowels per speaker[32]. The *iCAbS* (iterative cepstral analysis by synthesis) formant tracker (Clermont et al., 2007) was used to automatically extract and measure F1-F4 every 5 msec. This yielded between 10,000 and 30,000 F1-F4 measurements per speaker. LRs for LTFD1-4 were calculated using UBM-GMM in the same way as that detailed in Becker et al. (2008). In total there were 200 same speaker (SS) comparisons and 9800 different-speaker (DS) comparisons, as each recording for a speaker was divided in half. French et al. (2012) found LTFD1-4 to perform very well, with 97.4% of DS comparisons and 94% of SS comparisons identified correctly. In comparison with the discrimination levels of the MFCCs and VQ on the same data set, LTFD achieved similar error rates to the other methods; one method did not significantly outperform another. In terms of correlations between LTFD, MFCCs, and VQ, there were correlations found between the LRs produced from MFCCs and LRs calculated from the UBM-GMM analysis of LTFDs (r = 0.39). There was a weak correlation identified between VQ and LTFD globally (r = 0.12). However, there were some specific aspects of VQ that were more closely correlated with single LTFD measurements (e.g. raised larynx and LTFD1, r = 0.40). In conclusion, French et al. (2012) suggest the use of a vocal

[32] This was done using *Synthesis Toolkit CV* software that was adapted by Philip Harrison from J P French Associates.

tract output measurement (e.g. MFCC, VQ, LTFD) to be used as one of many tools in forensic speaker comparisons in order to "examine speech as varying human behavior."

The research in the studies presented above was carried out using a traditional (acoustic) phonetic approach, or an MFCC-based one. The discriminant performance of LTFD was tested in previous studies only using the GMM-UBM LR framework. At present, no previous studies have provided population statistics for LTFD in English or considered LTFD under an MVKD LR framework. This chapter will address both gaps.

## 5.3 Population Statistics for LTFD1-4

The following section discusses the collection of population statistics for LTFD in a large, linguistically-homogeneous group of 100 male speakers. These data serve as the first of their kind in providing detailed information on the distribution and variation that occurs in LTFD for a large group of individuals who speak Southern Standard British English (SSBE).

### 5.3.1 Methodology

Spontaneous speech recordings of 100 male speakers of SSBE, aged 18-25, were analyzed. The recordings were from Task 2 (a conversation between the speaker and his accomplice) of the DyViS database (Nolan et al., 2009). The recordings were automatically segmented to obtain a minimum of 50 seconds of concatenated vowels per speaker, and the iCAbS formant tracker (Clermont et al., 2007) was used to automatically extract and measure F1-F4 every 5 msec.

Population statistics were calculated by averaging all measurements for each formant for a single speaker and also taking the SD for each of those formants.

### 5.3.2 Results

The following section analyzes LTFD1-4 individually. The distributions for LTFD1 means and SDs are provided in Figures 5.1 and 5.2. The y-axis represents the number of speakers with mean LTFD formant frequencies that fall within a given range and the x-axis represents 10Hz-wide formant frequency bins, presented in Hertz (Hz).

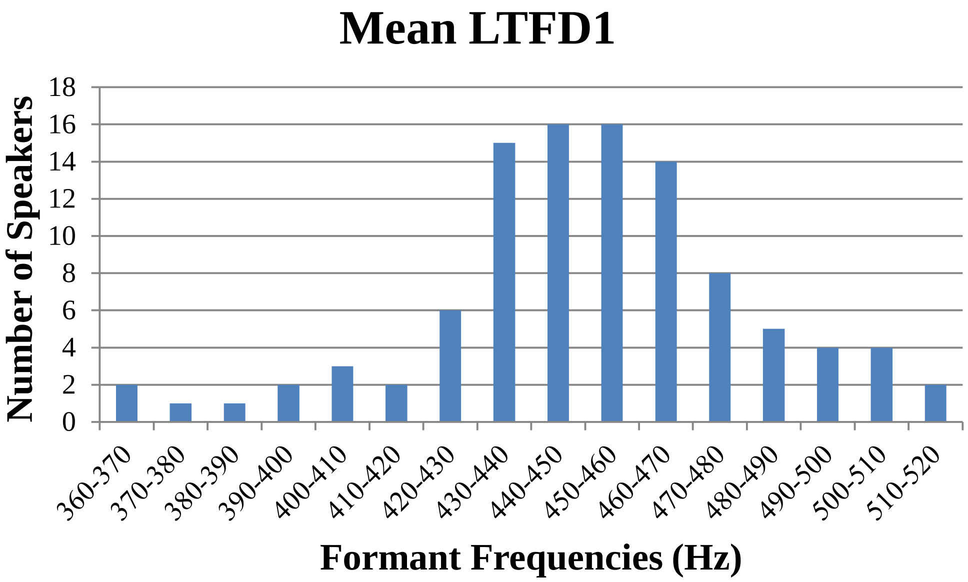


**Figure 5.1:** Distribution of mean LTFD1

Figure 5.1 shows a normal distribution with a slight negative skew due to suspected outliers (± 1.5 times the interquartile range) at 364.7Hz, 367.1Hz, 375.6Hz, 386.7Hz and one suspected outlier at 515.6Hz. The overall mean for the group LTFD1 is 451Hz, with a range of 364.7Hz to 515.6Hz. The SD of the means is 29.9Hz, and all 100 speakers' SDs fall within a 150.9Hz range.

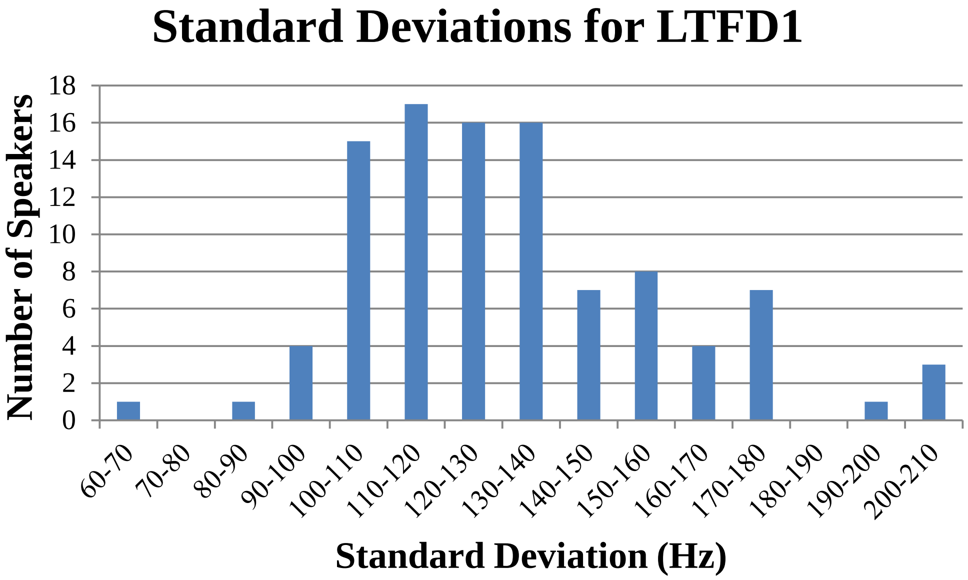


**Figure 5.2:** Distribution of standard deviation in LTFD1

The standard deviation values for LTFD1 within speakers follows a roughly normal distribution. There are three suspected outliers at 201.6Hz, 203.6Hz, and 209.8Hz. The mean SD is 131.4Hz, with a range of 64.8Hz to 209.8Hz. The SD of the mean SDs is 26.8Hz. All 100 speakers have SDs within 145Hz, which is a larger range (by 58.7Hz) than the range of means found in Figure 5.1.

The cumulative distribution graphs of LTFD1 means and SDs in Figure 5.3 and Figure 5.4, respectively, show the percentile within which a given LTFD1 mean or SD falls within the population. The y-axis is the cumulative proportion, and the x-axis represents formant frequencies in Hz.

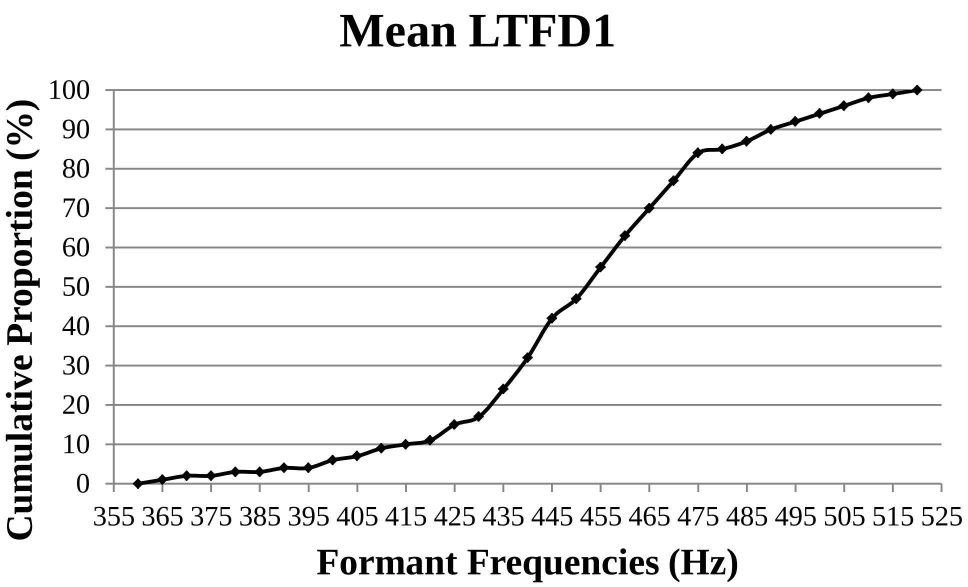


**Figure 5.3:** Cumulative percentages for mean LTFD1

The curve in Figure 5.3 is characterized by steepness in the central portion and gentle gradients in the first and third portions. Despite the steepness of the central section, the curve is overall a lot more gradient than was seen for AR in Chapter 4. ±1 SD from the mean gives a range between 421.1Hz and 480.9Hz, into which roughly 83% of the population tested here falls. The cumulative distribution of individual SDs is illustrated in Figure 5.4.

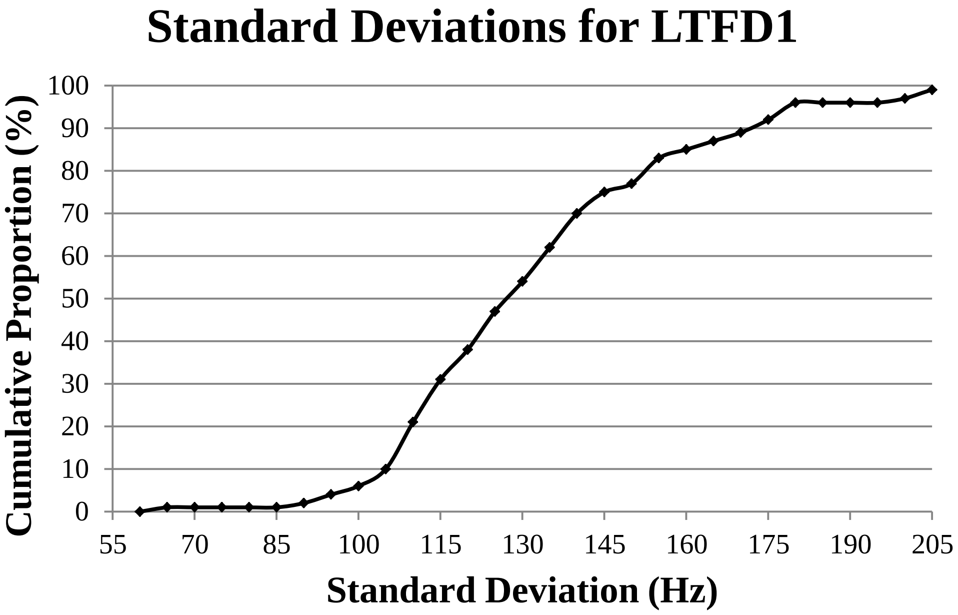


**Figure 5.4:** Cumulative percentages for standard deviation of LTFD1

The curve in Figure 5.4 is slightly steeper than that seen in Figure 5.3, and the beginning and end portions are less gradient. ±1 SD from the mean SD gives a 53.6Hz range between 104.6Hz and 158.2Hz, into which roughly 77% of the population falls. Given that the mean LTFD values are representative of the variation that occurs between speakers, and the SD values represent within-speaker variation, a variance ratio (Rose et al. 2006) can be calculated to ascertain which variation is higher. The LTFD1 variance ratio is 0.05, which is indicative of higher inter-speaker variation than intra-speaker variation.

The results for LTFD2 are presented in Figures 5.5 and 5.6. The graphs illustrate the population distributions for LTFD2 mean and SD.

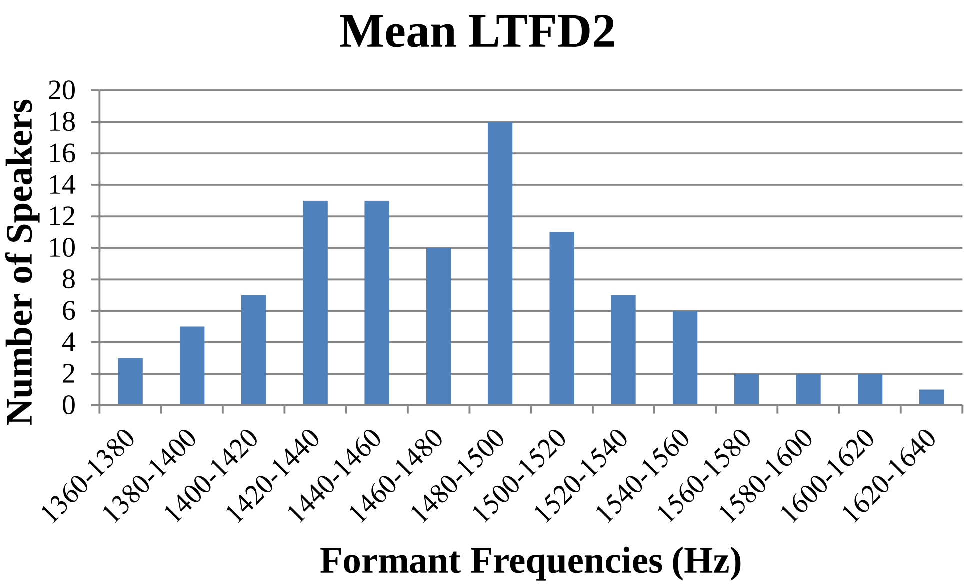


**Figure 5.5:** Distribution of mean LTFD2

Figure 5.5 has as a normal distribution with a slight positive skew due in part to a suspected outlier at 1633Hz. The overall mean LTFD2 for the group is 1476.7Hz, with a range of 1363.9Hz to 1633Hz. The SD of the LTFD2 means is 55.9Hz, and all 100 speakers fall within a 269.1Hz window.

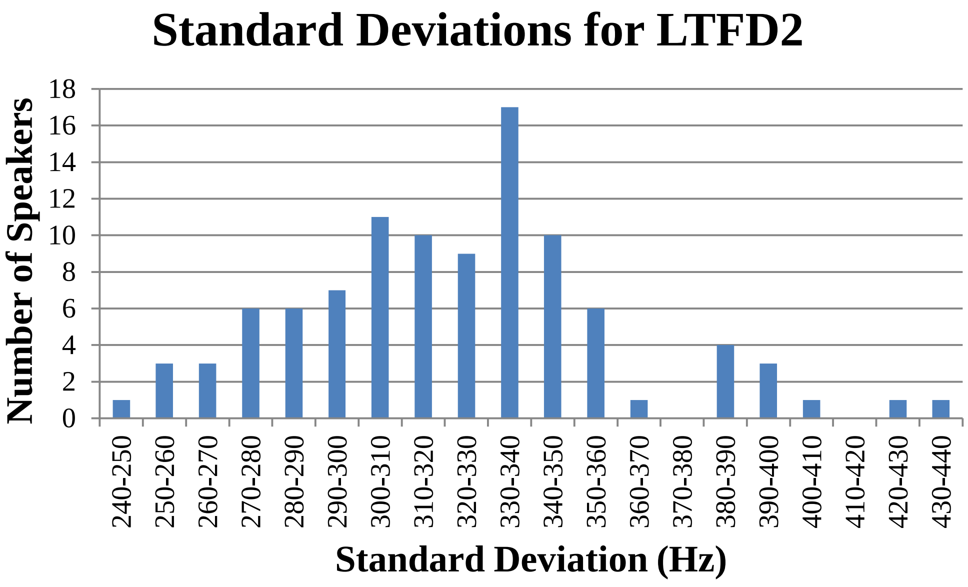


**Figure 5.6:** Distribution of standard deviation in LTFD2

The standard deviation for LTFD2 within speakers is again roughly normally distributed. There are two suspected outliers at 425.4Hz and 437.7Hz. The LTFD2 mean SD is 322.7Hz, with a range of 249.3Hz to 437.7Hz. The SD of the mean SDs is 37.7Hz. All 100 speakers have SDs within 188.4Hz.

The cumulative distribution graphs of means and SDs in Figure 5.7 and Figure 5.8, respectively, show the percentiles at which a given LTFD2 mean or SD falls within the population.

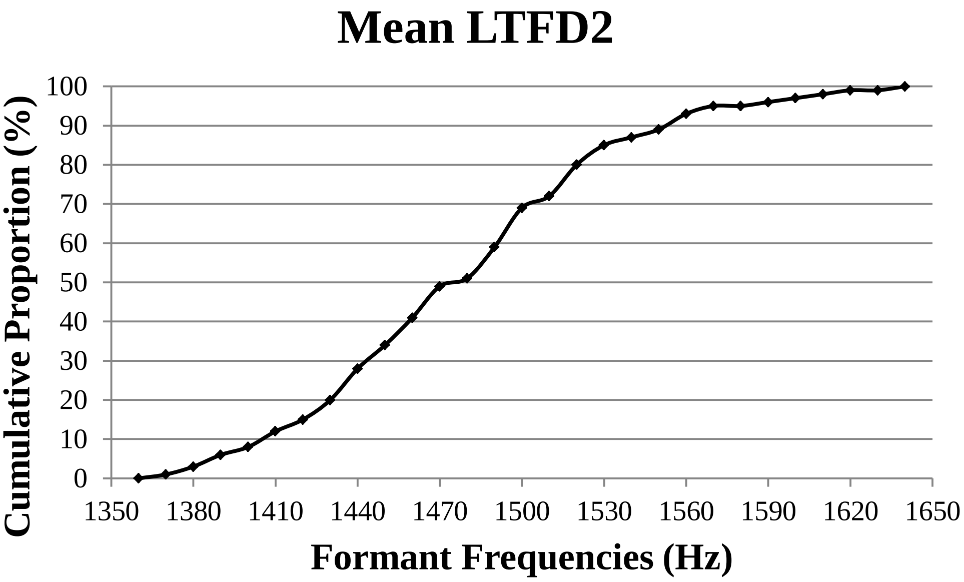


**Figure 5.7:** Cumulative percentages for mean LTFD2

The curve in Figure 5.7 is rather gradual compared to those in Figures 5.3 and 5.4, which is indicative of the relatively platykurtic distribution seen in Figure 5.5. ±1 SD from the mean gives a range between 1420.8Hz and 1532.6Hz, into which roughly 72% of the sample population falls. The cumulative distribution of individual LTFD2 SDs is illustrated in Figure 5.8.

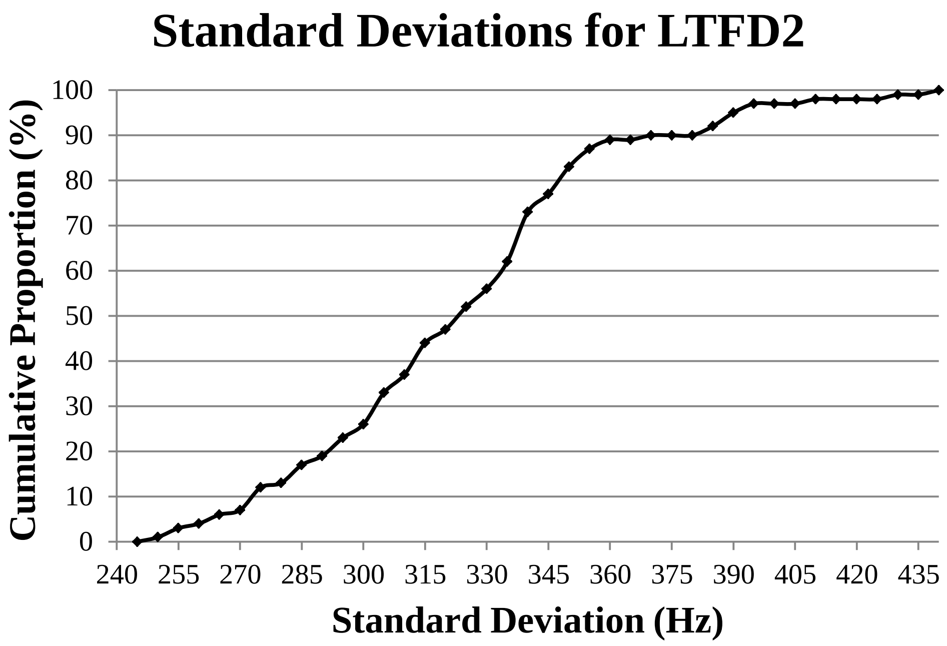


**Figure 5.8:** Cumulative percentages for standard deviation in LTFD2

The curve in Figure 5.8 is similar to that in Figure 5.7; however, the rate of increase of the middle portion in Figure 5.8 is much more variable. ±1 SD from the mean SD gives a range between 285Hz and 360.4Hz, into which roughly 71% of the sample population falls. Comparing the intra-speaker variation to the inter-speaker variation for LTFD2, there is a variance ratio 0.03. This indicates higher levels of variation within speakers than between them.

The results for LTFD3 are presented in Figures 5.9 and 5.10 below. The graphs represent the population distributions for LTFD3 mean and SD, respectively.

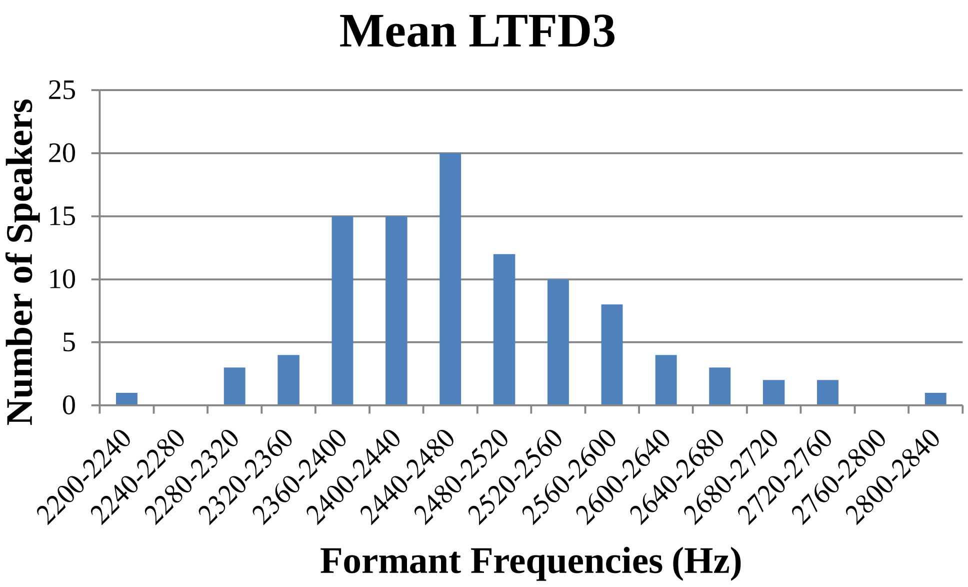


**Figure 5.9:** Distribution of mean LTFD3

Figure 5.9 has as a normal distribution with a slight negative skew. This is potentially due in part to a suspected outlier at 2824.4Hz. The overall mean LTFD3 for the group is 2478.5Hz, with a range of 2212.6Hz to 2824.4Hz. The SD of the LTFD3 means is 106.5Hz, and all 100 speakers fall within a 611.8Hz window.

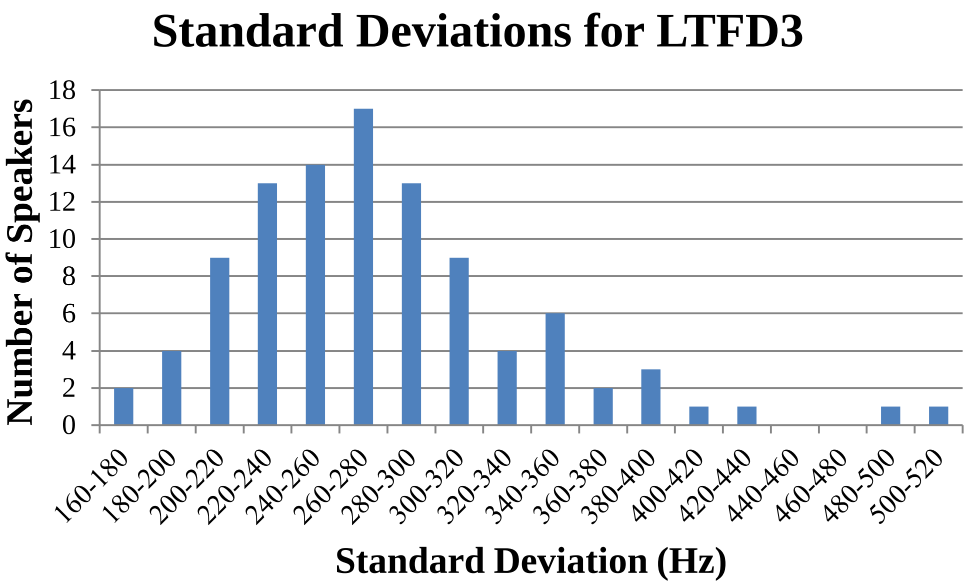


**Figure 5.10:** Distribution of standard deviation distribution in LTFD3

The standard deviation for LTFD3 within speakers is roughly normally distributed with a slight negative skew. It can be expected that the skew is due to the extreme outlier at 516Hz and three suspected outliers at 416.3Hz, 422.1Hz, and 491.9Hz. The mean SD is 277.9Hz, with a range of 168.2Hz to 516Hz. The SD of the LTFD3 SDs is 62.8Hz. All 100 speakers have SDs within 188.4Hz of each other.

The cumulative distribution graphs of LTFD3 means and SDs in Figure 5.11 and Figure 5.12, respectively, illustrate the percentile at which a given LTFD3 mean or SD falls within the population.

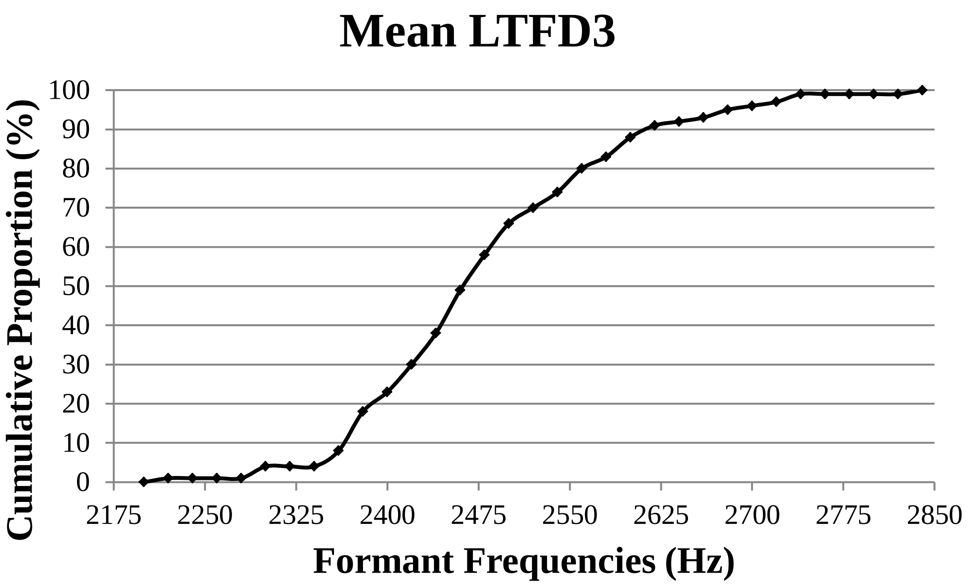


**Figure 5.11:** Cumulative percentages for mean LTFD3

The curve in Figure 5.11 is rather steep in the middle portion, with the beginning and end portions of the slope being more gradient. ±1 SD from the mean gives a range between 2372Hz and 2585Hz, into which roughly 70% of the population falls. The cumulative distribution of individual speakers' SDs for LTFD3 is presented in Figure 5.12.

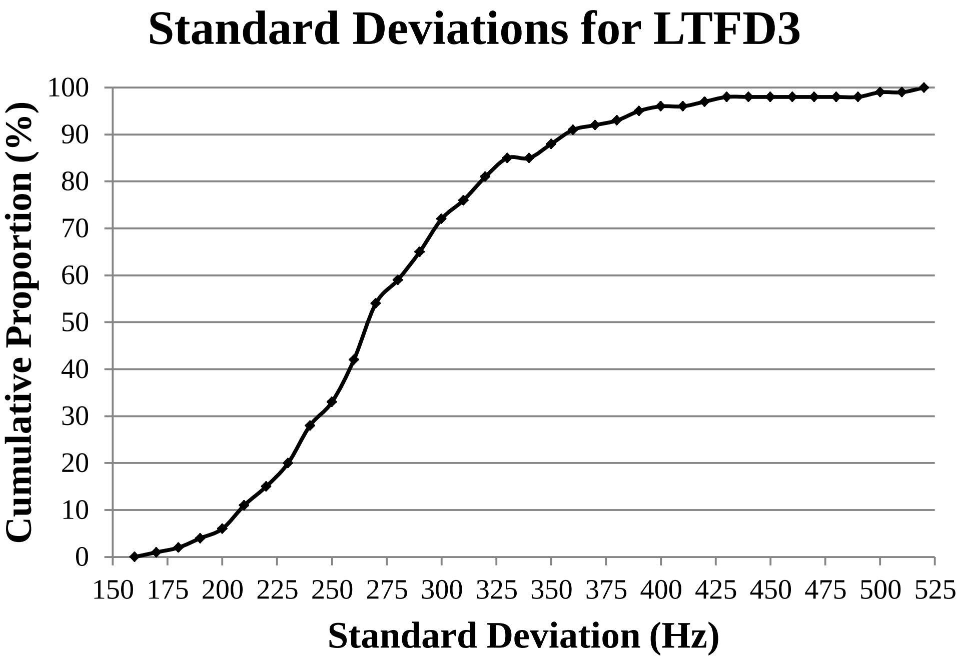


**Figure 5.12:** Cumulative percentages for standard deviation in LTFD3

The curve in Figure 5.11 is rather steep, with the end portion of the slope, which starts around 375Hz, being more gradient than the beginning. ±1 SD from the mean gives a range between 215.1Hz and 340.7Hz, into which roughly 68% of the sample population falls. The calculated variance ratio for LTFD3 is 0.15. This suggests it is more likely that one will find higher levels of variation within a speaker than between that speaker and the rest of the population, which was also the case for LTFD1 and LTFD2.

The results for LTFD4 are presented in Figures 5.13 and 5.14 below. The graphs display the population distributions for LTFD4 mean and SD, respectively.

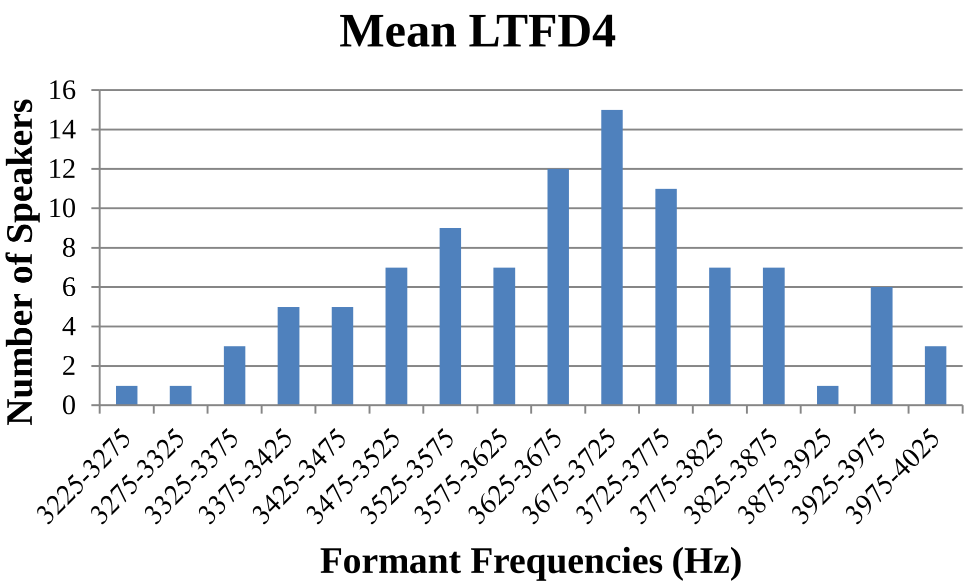


**Figure 5.13:** Distribution of mean LTFD4

Figure 5.13 shows that mean LTFD4 has as a fairly normal distribution, though with a slight negative skew. However, there are no suspected (1.5 x the interquartile range) or definite (3 x the interquartile range) outliers in the data. This could simply be the natural distribution of the data or perhaps it is revealing measurement errors that occurred for those individuals that appear to have lower LTFD4 means. The overall mean LTFD4 for the group is 3660.9Hz, with a range of 3249.9Hz to 4019.5Hz. The SD of the LTFD4 means is 170.9Hz, and all 100 speakers fall within a 769.6Hz window.

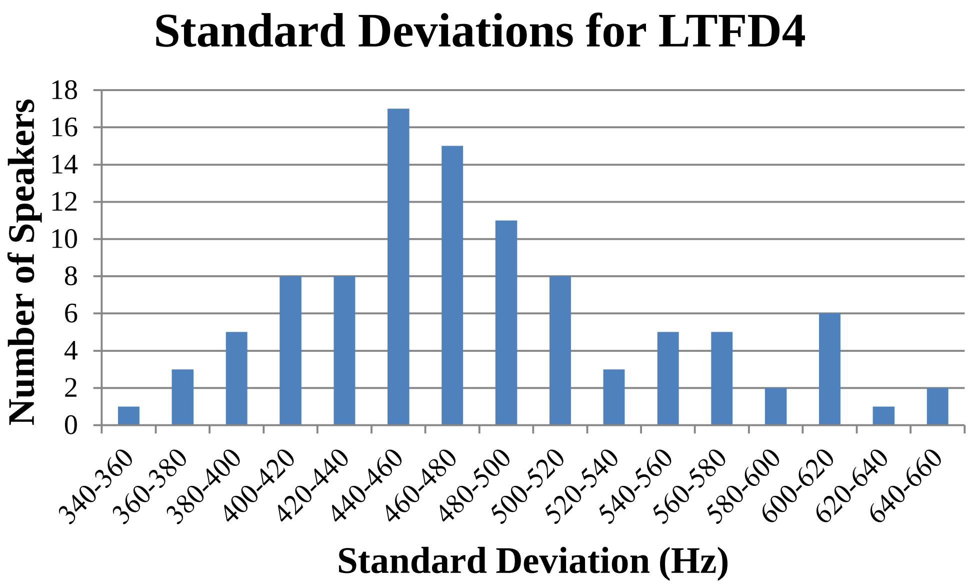


**Figure 5.14:** Distribution of standard deviation in LTFD4

The standard deviation for LTFD4 within speakers is roughly normally distributed. There are three suspected outliers at 633.3Hz, 640.5Hz, and 649.5Hz. The mean SD is 482.2Hz, with a range of 356.8Hz to 649.5Hz. The SD of the LTFD4 SDs is 67.2Hz. All 100 SSBE speakers have SDs within 292.7Hz of each other.

The cumulative distribution graphs of LTFD4 means and SDs in Figure 5.15 and Figure 5.16, respectively, show the percentile at which a given LTFD4 mean or SD falls within the population.

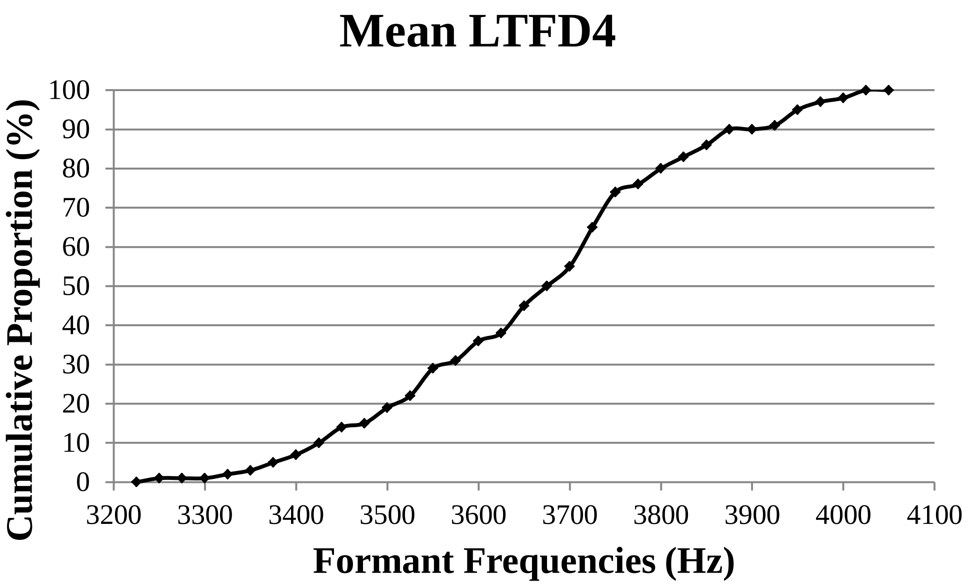


**Figure 5.15:** Cumulative percentages for mean LTFD4

The curve in Figure 5.15 is characterized by a rather gradual increase, which is more similar to the curve in Figure 5.7 than it is to the curves illustrated in Figures 5.3 and 5.11. ±1 SD from the mean gives a range between 3490Hz and 3831.8Hz, in which roughly 63% of the sample population falls. The cumulative distribution of individual SDs for LTFD4 is illustrated in Figure 5.16.

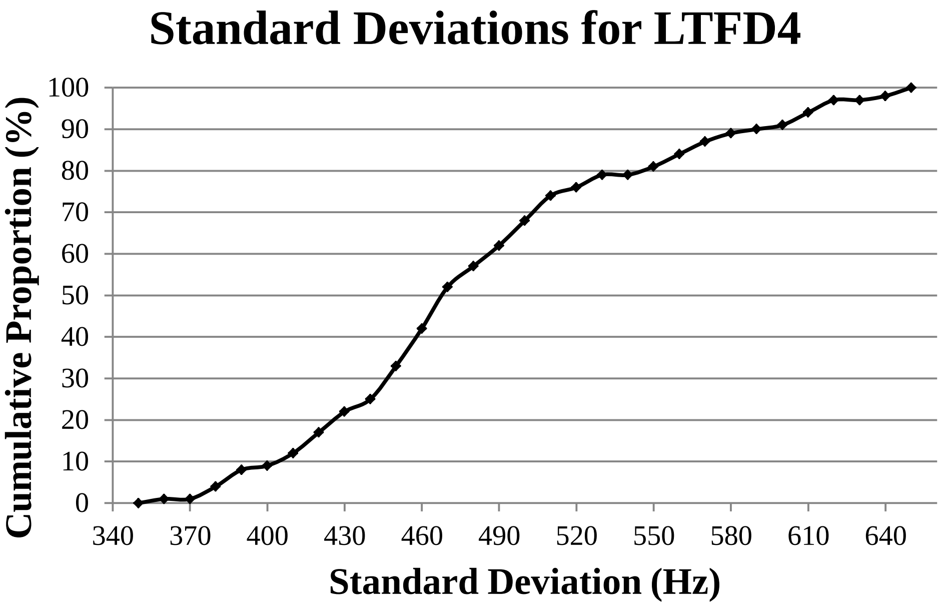


**Figure 5.16:** Cumulative percentages for standard deviation in LTFD4

The curve in Figure 5.16 has a rather gradual, almost linear slope. The graph has a similar shape to the curve shown in Figure 5.15. ±1 SD from the mean SD gives a range between 415Hz and 549.4Hz, into which roughly 65% of the population falls. The variance ratio for LTFD4 is 0.13, which again indicates that one will be more likely to find higher levels of variation within a speaker than between that speaker and the rest of the population, which we also saw in the case of LTFD1-3. Given that variance ratios greater than one indicate that more variation occurs within individuals than between them, the LTFDs with higher variance ratios discriminate better between individuals than do those with lower variance ratios. LTFD3 had the highest ratio at 0.15, followed by LTFD4 at 0.13, LTFD1 at 0.05, and finally LTFD2 at 0.03. The LR results in the next section (§ 5.4) are hypothesized to follow these predictions.

### 5.3.3 LTFD1-4 Results Compiled

For simplification purposes the overall LTFD population results are displayed in Table 5.1 and 5.2. Table 5.1 compiles LTFD1-4 results for between-speaker variation, while Table 5.2 presents the compilation of LTFD1-4 results for within-speaker variation. The first columns identify the formant, and the second through fourth columns contain mean SD, range (of SD), and SD (of SD).

**Table 5.1:** Overall between-speaker results for LTFD1-4

| LTFD | Mean (Hz) | Range (Hz) | SD (Hz) |
|---|---|---|---|
| 1 | 451.0 | 364.7-515.6 | 29.9 |
| 2 | 1476.7 | 1369.9-1633.0 | 55.9 |
| 3 | 2478.5 | 2212.6-2824.4 | 106.5 |
| 4 | 3660.9 | 3249.9-4019.5 | 170.9 |

**Table 5.2:** Overall within-speaker results for LTFD1-4

| LTFD | Mean SD (Hz) | Range of SD (Hz) | SD of SD(Hz) |
|---|---|---|---|
| 1 | 131.4 | 64.8-209.8 | 26.8 |
| 2 | 322.7 | 249.3-437.7 | 37.7 |
| 3 | 277.9 | 168.2-516.0 | 62.8 |
| 4 | 482.2 | 356.8-649.5 | 67.2 |

The mean formant measurements for LTFD1-4 in Table 5.1 are very similar to those of [ə], where F1 is about 500 Hz, F2 is 1500 Hz, F3 is 2500 Hz, and F4 is 3500 Hz (Johnson, 2003). Given that LTFD is an average across all vowel phonemes, some type of central (with respect to the vowel space) vowel would be expected. The results in Table 5.1 and 5.2 also show that LTFD3 and LTFD4 have the smallest ratios of mean SD to mean formant value (277.9: 2478.5 and 482.2: 3660.9). This is indicative of the two higher formants being more stable within speakers, suggesting that they will be better speaker discriminants than the lower formants (this also coincides with the variance ratios from § 5.3.2 above).

## 5.4 Likelihood Ratios for LTFD

The discriminant results of LTFD are presented in the following section. In the first part, LR results are provided for individual formants as well as in combination with other formants. The second part considers the effects that ‘package length’ (Moos, 2010) has on LR results for LTFD.

### 5.4.1 Methodology

Likelihood ratios (LRs) were computed using a MatLab implementation of Aitken and Lucy’s (2004) Multivariate Kernel-Density formula (Morrison, 2007) for the 100 male speakers in DyViS Task 2.  The MVKD formula was originally developed for use with evidence that included repeated measures of a given parameter (Aitken and Lucy, 2004). However, LTFD considers evidence from all possible vowel categories, resulting in raw data that can be extremely varied. For this reason, the raw formant data were averaged over 0.5 sec windows (a total of 100 raw data measurements per formant constituted a single token) for F1-F4 in order to obtain what Moos refers to as “packages” (Moos, 2010). There were 100 to 284 (LTFD1-4) measurements per speaker, with a mode of 100 tokens. An intrinsic discrimination method was used to calculate LRs, whereby speakers 1-50 acted as the test set and speakers 51-100 acted as the reference set. LRs are calculated for LTFD1-4 individually as well as in combinations relevant to forensic casework.

These combinations were chosen with respect to common practices in the field (§ 3.9.1.1). Traditionally, the two formants most commonly used in casework and sociolinguistic studies are F1 and F2, which are measured in order to reveal aspects of an individuals’ vowel space (Ash, 1988; Milroy and

Gordon, 2003). Some FSS experts still analyze only these two formants (§ 3.9.1.1), and therefore LTFD1 and LTFD2 are considered in combination.

LTFD research has reported that formants are often prone to variation. It is common for a case to involve the comparison of material from a telephone recording (cellular phone or landline) against a directly-recorded sample (often a police interview). Due to the limited bandwidth of transmission over the telephone (the 340-3700Hz band) there are many acoustic properties of the signal that are often affected (Foulkes and French, 2012). The most notable are an artificial increase in F1 values and formants close to 3700Hz disappearing. Often F4 is missing from the signal altogether (Künzel, 2001; Byrne and Foulkes, 2004). A similar effect has also been reported for recordings made using the video and voice recorders in cellular phones (Gold, 2009). For this reason, some experts avoid F1 and F4 altogether, meaning that only LTFD2 and LTFD3 are analyzed. In addition, it is important to note that analysts must be aware that the distance between the microphone (of the recording device) and the talker (in conjunction with the room acoustics) can also have effects on formant measurements (Vermeulen, 2009).

A majority of experts (63%) reported measuring F1-F3 in casework (§ 3.9.1.1), and therefore LTFD1-LTFD3 are considered, as they are the most commonly-analyzed combination of formants. Finally, LTFD1-4 are considered in combination to represent the ideal case where F1-F4 are all measureable. This also provides the upper boundary in terms of the maximum number of features within a parameter that can be used (for the given data) to achieve the best possible performance. All LR results are considered in terms of system performance (EER and Cllr) and the magnitude of strength of evidence (Champod and Evett, 2000).

### 5.4.2 Results for LTFD1-4: Individually and in Combination

The results for the calculation of LRs on LTFD1-4 individually are summarized in Table 5.3. The leftmost column represents the LTFD that was analyzed and whether it relates to same-speaker (SS) or different-speaker (DS) comparisons. The second column indicates the percentage of SS or DS comparisons that were correctly classified. Correct SS comparisons have a log likelihood ratio (LLR) above zero, and correct DS comparisons have an LLR of less than zero. The mean LLR is in the third column, followed by the minimum and maximum LLR in the next two columns. The final two columns present the EER and Cllr values.

**Table 5.3:** Summary of LR-based discrimination for LTFD1-4 (100 speakers)

| Comparisons | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **LTFD1 SS** | 72.0 | 0.224 | -2.158 | 1.902 | .2806 | .8840 |
| **LTFD1 DS** | 71.7 | -4.858 | -68.768 | 1.993 | | |
| **LTFD2 SS** | 70.0 | 0.162 | -1.077 | 1.259 | .3165 | .8119 |
| **LTFD2 DS** | 67.5 | -1.939 | -27.814 | 1.602 | | |
| **LTFD3 SS** | 88.0 | 0.288 | -8.373 | 3.743 | .1700 | 1.0731 |
| **LTFD3 DS** | 80.6 | -11.857 | -139.273 | 1.734 | | |
| **LTFD4 SS** | 68.0 | 0.238 | -2.258 | 1.378 | .2214 | .8085 |
| **LTFD4 DS** | 80.2 | -11.574 | -124.808 | 1.301 | | |

Table 5.3 shows that, overall, LTFD3 has the lowest EER (.1700) but the highest Cllr (1.0731). LTFD4 performed second-best in terms of EER (.2214) and best for Cllr (.8085). The highest EER was for LTFD2 at .3165, but it had the second-lowest Cllr (.8119). Overall, SS comparisons achieved a higher proportion of correct results than did DS comparisons, with the exception of LTFD4, where DS

comparisons performed 12.2% better. The strength of evidence (i.e. mean LLR) is stronger for DS comparisons than SS comparisons; mean LLRs are between -11.857 and -1.939. For SS comparisons, the magnitude of the strength of evidence is lower, ranging from 0.162 to 0.288. With respect to Champod and Evett's verbal scale (2000, p. 240) the LLR scores for SS would not even constitute limited evidence in support of the prosecution hypothesis. However, there are some cases where a formant individually achieves a stronger strength of evidence, as per LTFD3 with its maximum LLR of 3.743 (moderately strong evidence).

Examining LTFD1-4 individually, the results in Table 5.3 suggest that LTFD3 performs the best overall, followed by LTFD4, LTFD1, and finally LTFD2. Despite returning the highest Cllr, LTFD3 has the highest percentage of correct SS and DS comparisons. It also offers the lowest EER, while providing the strongest strength of evidence for SS and DS. The suspected reason for Cllr being at its highest for LTFD3 is that Cllr appears to be greatly affected by parameters (e.g. vowels) that produce wider ranges and higher magnitudes of LLR. While producing these correct SS and DS comparisons with significant strengths of evidence, it also tends to cause comparisons to yield incorrect SS and DS comparisons with high strengths of evidence. For this reason, high Cllrs appear to be being calculated for parameters that have the potential to offer more in terms of correctness and the magnitude of the strength of evidence. This was seen in Chapter 4, where although AR performed very poorly as a discriminant, it achieved a lower Cllr than LTFD3, because the magnitude of the

AR LLRs overall were smaller. The same is true of LTFD2, which has the lowest Cllr and also the smallest magnitude LLRs[33].

Individually, LTFD1-4 performed relatively well, but the combination of the formants can potentially yield even better performances. Table 5.4 below follows the same structure as Table 5.3.

**Table 5.4:** Summary of LR-based discrimination for different LTFD1-4 combinations (100 speakers)

| Comparisons | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **LTFD1+2 SS** | 70.0 | 0.417 | -2.472 | 2.761 | .2041 | .7648 |
| **LTFD1+2 DS** | 85.0 | -7.477 | -76.391 | 1.996 | | |
| **LTFD2+3 SS** | 76.0 | 0.334 | -7.828 | 3.768 | .1392 | .9630 |
| **LTFD2+3 DS** | 89.9 | -14.173 | -156.130 | 1.956 | | |
| **LTFD1+2+3 SS** | 74.0 | 0.625 | -7.632 | 3.676 | .1147 | 1.0161 |
| **LTFD1+2+3 DS** | 94.3 | -19.307 | -155.807 | 3.007 | | |
| **LTFD1+2+3+4 SS** | 84.0 | 1.160 | -5.292 | 5.466 | .0414 | .5411 |
| **LTFD1+2+3+4 DS** | 97.4 | -29.228 | -162.931 | 2.854 | | |

Four different LTFD combination scenarios are presented in Table 5.4. LTFD1+2 performed the worst with respect to EER (.2041), but was the best in terms of Cllr (.7648). LTFD1+2+3+4 performed the best with respect to EER, which was .0414, and had the lowest Cllr (.5411). The highest proportion of correct SS and DS comparisons was also returned by LTFD1+2+3+4, with 84% and 97.4%, respectively. LTFD1+2 had the lowest proportion of correct SS and DS comparisons with 70% and 85%, respectively. LTFD1+2+3 performed better than LTFD2+3 with higher proportions of correct DS comparisons, mean LLR, and EER. Overall, the combination of LTFD1+2+3+4 outperformed the other three combinations as defined by EER and the proportions of correct SS and DS comparisons.

[33] An additional explanation for the poor Cllr values could be that the system is not optimally calibrated (see § 8.5.3) as was also seen in § 4.6.2.1.

The Tippett plots for the four LTFD1-4 combinations are presented in Figures 5.17-5.20 below.

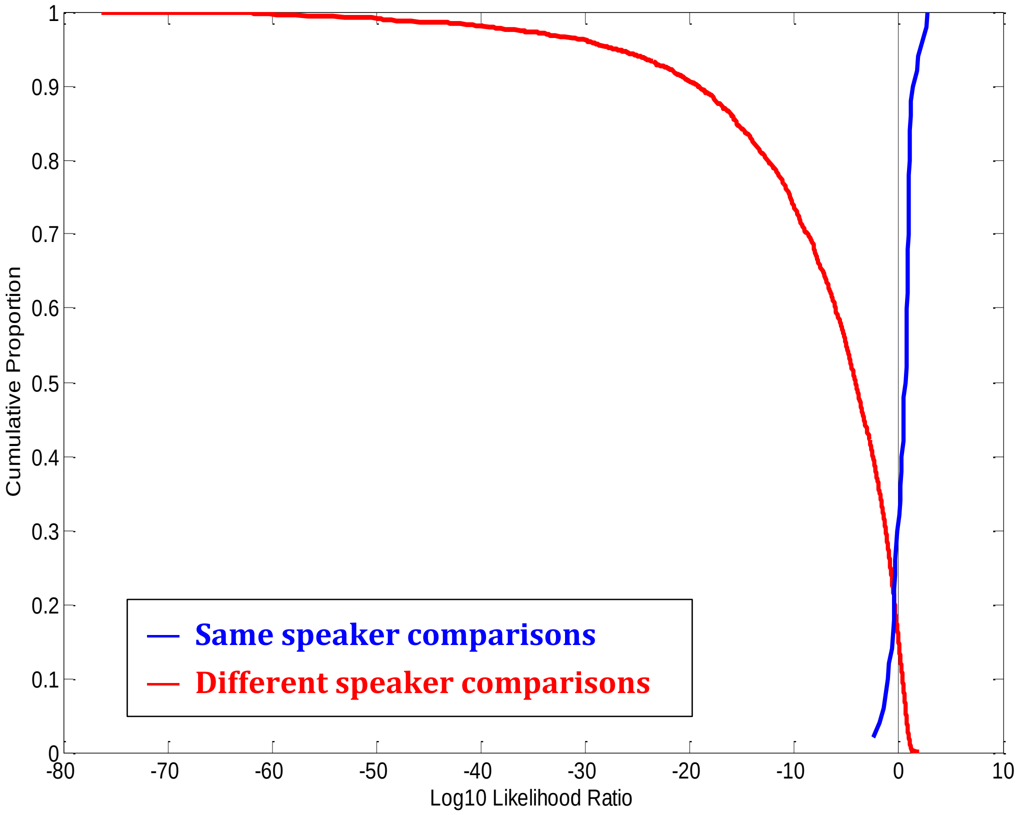


**Figure 5.17:** Tippett plot of LTFD1+2

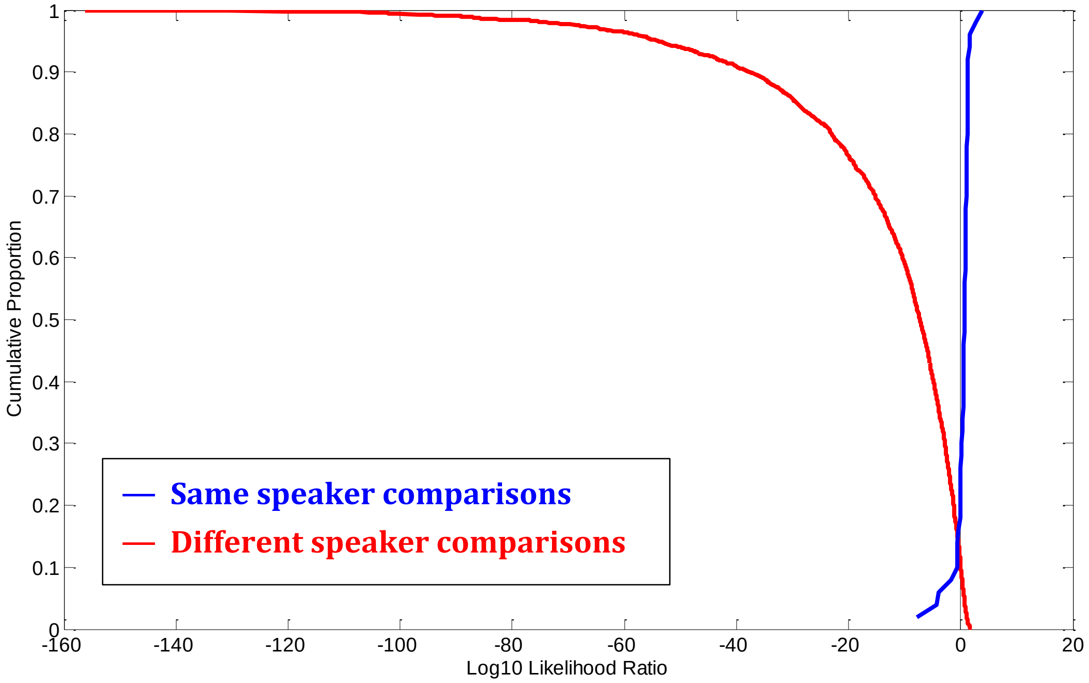


**Figure 5.18:** Tippett plot of LTFD2+3

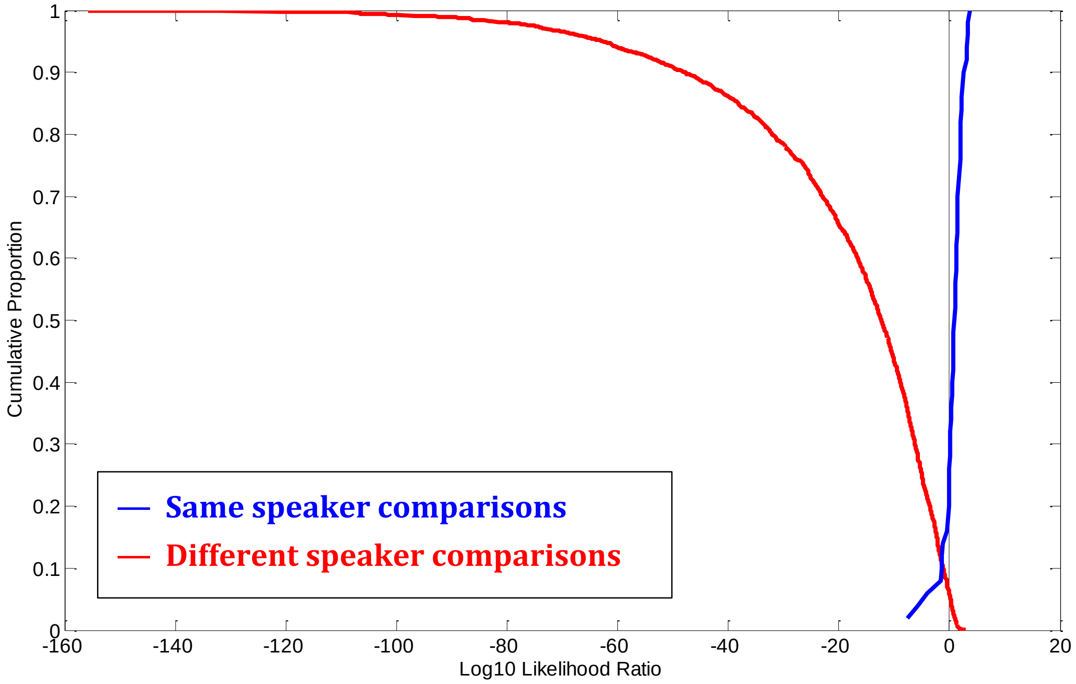


**Figure 5.19:** Tippett plot of LTFD1+2+3

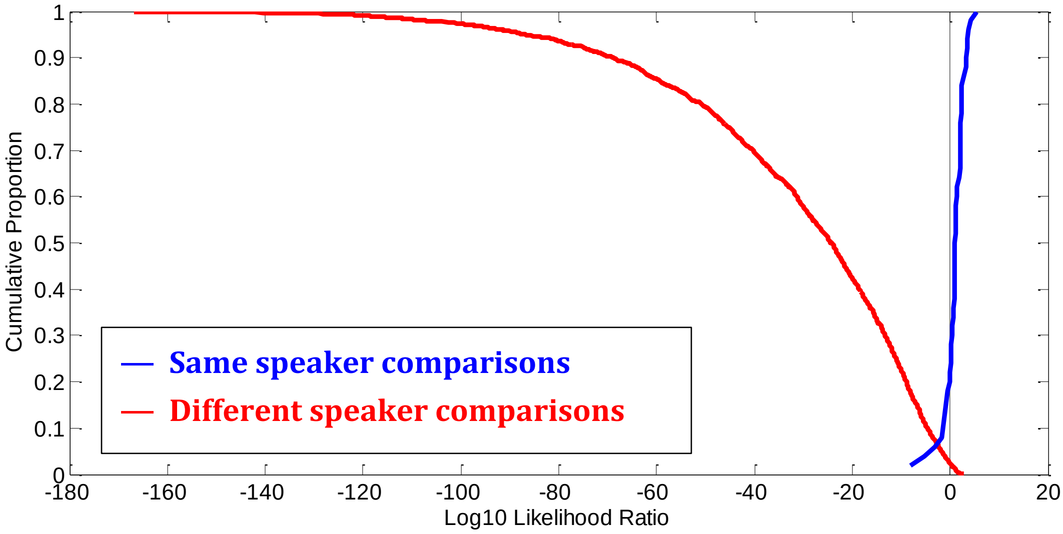


**Figure 5.20:** Tippett plot of LTFD1+2+3+4

Figures 5.17-5.20 illustrate the range of LLR values calculated for the four LTFD1-4 scenarios. LTFD1+2+3+4 had the largest positive values for SS LLR and the largest negative values for DS LLR (i.e. best strengths of evidence in both cases), and the best overall mean LLRs (SS = 1.16, DS = -29.228). LTFD1+2 had the smallest LLR ranges for both SS and DS, and the weakest mean DS LLR. LTFD2+3 yielded the weakest mean SS LLR, and the second-weakest mean DS LLR. LTFD1+2+3+4 offered the strongest LLR for same-speaker pairs at 5.466 (very strong evidence), and the strongest minimum LLR for DS at -162.931. Overall, LTFD1+2+3+4 was the best combination of formants, followed by LTFD1+2+3, LTFD2+3, and finally LTFD1+2. In comparison to the figures for LTFDs for individual formants, the four combination scenarios were able to significantly lower EER and improve the proportion of correct SS and DS comparisons.

### 5.4.3 Results of Package Length

LTFDs were measured every 5msec in the current study; however, to test the discriminant power of LTFD using MVKD[34], multiples of 5msec *packages*[35] were created from multiple LTFD measurements to create localized LTFD tokens (which are equivalent to short-duration portions of the recording). The length of the package over which a distribution is calculated can vary. A package length of 0.5 seconds was chosen for the study, as it was found to yield the lowest EERs. The effects of package length variation can be seen in Table 5.5 for the LTFD combination of LTFD1+2+3+4. The size of the package length is provided in the first column, followed by the percentage of SS and DS comparisons that were correct. The mean, minimum, and maximum LLRs for SS and DS comparisons are found in columns four through nine, while EER and Cllr are provided in the last two columns.

**Table 5.5:** Package length variability

| Package Length | SS % Correct | DS % Correct | Mean SS LLR | Mean DS LLR | Min SS LLR | Min DS LLR | Max SS LLR | Max DS LLR | EER | Cllr |
|---|---|---|---|---|---|---|---|---|---|---|
| ***.25 sec*** | 76 | 97.96 | 0.90 | -35.58 | -6.51 | -199.61 | 5.64 | 2.93 | 0.043 | 0.775 |
| ***.5 sec*** | 84 | 97.43 | 1.16 | -29.23 | -5.29 | -162.93 | 5.47 | 2.85 | 0.041 | 0.541 |
| ***1 sec*** | 88 | 96.73 | 1.34 | -24.17 | -4.18 | -134.33 | 5.29 | 2.79 | 0.042 | 0.400 |
| ***2.5 sec*** | 94 | 95.76 | 1.52 | -17.67 | -3.17 | -98.41 | 4.99 | 2.88 | 0.043 | 0.281 |
| ***5 sec*** | 96 | 94.82 | 1.60 | -13.85 | -2.45 | -84.60 | 4.77 | 2.90 | 0.042 | 0.239 |
| ***10 sec*** | 98 | 92.78 | 1.55 | -9.27 | -2.59 | -62.94 | 4.41 | 2.68 | 0.056 | 0.257 |

The results in Table 5.5 suggest that package length affects the discriminant performance of LTFD. An increase in package length corresponds to an improvement in correct SS comparisons and Cllr, while there is a decrease in

[34] LTFDs in their raw form readily lend themselves to a UBM-GMM algorithm for calculating LRs. However, in order to test the MVKD formula, LTFDs were put into packages, as the MVKD formula is not equipped to handle streams of data.

[35] Package length was also used by Moos (2010) to determine stability within LTFD over varying quantities of data. The packages are used in a similar way here, but are evaluated in terms of their validity.

correct DS comparisons. In general, EER improves with the decrease in package length; however, EER appears to have a threshold around 0.5 seconds, above which it no longer improves.

It is important to note that these results apply only to the given study, where the total duration of material per speaker was around 50 seconds. It could be the case that package length has different effects on speech samples longer or shorter than the 50 seconds used in the current study. However, this analysis serves as a starting point for further investigation into the effects of variability of package length and the overall length of speech samples on LTFD results.

## 5.5 Discussion

The following section considers the discriminant value of higher formants and lower formants, and also compares the results from the present study with those from previous studies that used GMM-UBM-based LR calculations (Becker et al., 2008; French et al. 2012).

### 5.5.1 Discriminant Value of Higher Formants

The results from individual LTFD LRs revealed that LTFD3 and LTFD4 performed better than the lower formants, LTFD1 and LTFD2, in discriminating between speakers. These results suggest that the higher formants carry more speaker-discriminatory information than the lower formants (also seen in Jessen, 1997; McDougall, 2004; Moos, 2010; Simpson, 2008; Clermont et al., 2008; Hughes, 2013). Table 5.6 provides an overview of previous studies investigating the discriminant ability of formants where F3 (a higher formant) was also found to outperform lower formants (F1 and F2).

**Table 5.6:** Overview of discriminant formant studies where F3 performs best

| Study | Data | Formants Considered | Measurements | Most Discriminant |
|---|---|---|---|---|
| Jessen (1997) | German; 20 speakers | F1-3 | Peaks in spectra | F3 |
| McDougall (2004) | Australian English; 5 speakers | F1-3 | Dynamic | F3 |
| Moos (2010) | German; 71 speakers | F1-3 | LTFD | F3 |
| Simpson (2008) and Clermont et al. (2008) | British English; 25 speakers | F1-3 | temporal midpoint of formant | F3 |
| Hughes (2013) | British English; 97 speakers | F1-3 | Dynamic | F3 |

The explanation for the better performance of higher formants than lower formants (as seen in this study, and those listed in Table 5.6) can be obtained by recourse to phonetic theory. The first and second formants are responsible for encoding phonetic content (Ladefoged, 2006), where (lower) frequencies are related in large part to tongue position: the first formant correlates inversely with tongue height and the second formant is associated with tongue frontness/backness (Clark and Yallop, 1990, p. 268). The range of F1 and F2 values a speaker produces will be relatively constrained by the size and shape of his/her vocal tract, while the given configuration of a speaker's vocal tract will determine its F1 and F2 values. In general, the lower formants (i.e. F1 and F2) do not encode speaker-specific information; rather, they are responsible for conveying phonetic content.

Contrastively, higher formants (specifically F3 and F4) have been identified as encoding speaker-specific information, which makes sense given that they are less affected by behavioral and physiological variation than are lower formants (McDougall, 2004). This is because F3 and F4 are associated with the resonances in the smaller cavities of the vocal tract, which allow for less intra-speaker variation (i.e. smaller cavities offer a smaller space in which resonances are produced; Peterson, 1959). However, the inter-speaker variation of F3 and F4 is limited with respect to variation in the size of the vocal tract (which does not show a wide range of variation; Xue and Hao, 2006). It is important to note that Stevens and French (2012) have shown F3 to be correlated in part to voice qualities that involve the backing of the tongue body, an articulatory setting which was adopted by the majority of speakers in the accent group they studied (SSBE speakers). The same was also found for speakers of American English, where post-vocalic rhoticity results in the lowering of F3 (Alwan et al., 1997). This means that although F3 is in part responsible for differences in voice quality, which is to a large extent speaker-specific, to some degree F3 can also encode accent information, specifically that associated with tongue-body orientation (e.g. retracted tongue-body and a pharyngealized voice quality; Laver, 1994).

To this extent, the suggestion that higher formants carry more speaker-discriminant information than lower ones is borne out in the current research, and provides an argument in support of the good performance of LTFD3 and LTFD4 in the present study.

### 5.5.2 Comparison of LTFD, MFCC, MVKD, and GMM-UBM Results

The results presented in the current study were calculated using the MVKD formula. However, GMM-UBM has also been used on the same data (French et al. 2012), and LTFD on German data (Becker et al., 2008). The MFCC results (French et al., 2012) and the LR results from French et al. (2012) and Becker et al. (2008) for LTFD are compared (Table 5.7) to the results found in the present study.

**Table 5.7:** Summary of LR-based discrimination for LTFD and MFCC in the current study and competing studies

| | LTFD1+LTFD2+LTFD3 | | | LTFD1+LTFD2+LTFD3+LTFD4 | | | MFCC | | |
|---|---|---|---|---|---|---|---|---|---|
| | *SS* | *DS* | *EER* | *SS* | *DS* | *EER* | *SS* | *DS* | *EER* |
| **Current study** | 74% | 94.3% | .1147 | 84% | 97.4% | .0414 | - | - | - |
| **French et al. (2012)** | - | - | - | 94% | 97.4% | - | 100% | 95% | - |
| **Becker et al. (2008)** | - | - | .053 | - | - | - | - | - | - |

The LTFD results from all three studies are generally similar[36] regardless of whether GMM-UBM or MVKD was used. However, given that French et al. (2012) and the current study are based on the same recordings, it would be expected that SS comparison results were more similar than they are (94% to 84%, respectively). This could suggest that for LTFD it is preferable to use GMM-UBM over MVKD. However, it is important to note that 100 SS comparisons were made by French et al. (2012), whereas the current study only conducted 50 SS comparisons. Therefore, it is plausible that this 10% difference could be due in part to the disparity in sample size (10% is equivalent to five SS comparisons).

The tendency (albeit a small one) is for LTFD to miss SS pairs, and for MFCC to mistake DS pairs for SS pairs. In view of this, it could be argued that in the context of security, where investigators are working to put together a list of potential suspects, MFCCs would be the preferred analysis. This is because MFCCs are more likely to include additional suspects (despite their innocence) rather than miss them entirely. Additionally, MFCCs tend to over-estimate

[36] Becker et al. (2008) also included results using bandwidths. However, those results are not presented here.

similarity when comparing non-similar speaker pairings. In a judicial context, the opposite could be argued: LTFD analysis should be preferred, insofar as it is less likely to result in innocent suspects being misidentified as criminals. This is because LTFD has the tendency to be under-sensitive when it comes to identifying a guilty suspect as the criminal (by making judgments of non-similarity when the samples are in fact similar by virtue of having been spoken by the same talker).

## 5.6 Conclusion

Overall, the results presented in this chapter suggest that LTFD is a good speaker discriminant, despite all LTFDs having variance ratios that imply intra-speaker variation that is higher than inter-speaker variation. The combinations of LTFD1-4 in § 5.4.2 achieved higher levels of discrimination than single LTFDs. The best combination, LTFD1+2+3+4, had an EER of only 0.0414, which is extremely low compared to those found for AR in Chapter 4. Following the results of the survey reported in Chapter 3, it appears that experts were correct in identifying formants (in one form or another; e.g. LTFD, or for individual phonemes) as one of the most useful speaker discriminants.

A known limitation of LTFD results from the study by Moos (2010), where the values of LTFD means were higher in read speech than spontaneous speech. It appears that speaking style can have a large impact on LTFD results. It is important to consider in casework whether there is enough material available to work with, and whether the material in the suspect and criminal recordings is comparable, before carrying out an LTFD analysis.

The most attractive aspect of LTFD may not be in its successful results, but in the fact that LTFD is not correlated with a number of other parameters (Moos 2010). Correlation is often a challenge when vowels are analyzed individually and later there is a desire to combine results from those multiple vowels. This often results in a scenario where certain vowels are inevitably correlated, and following "naïve Bayes" (the combination of evidence through the multiplication of individual LRs only when pieces of evidence are mutually exclusive; Kononenko, 1990) they cannot be considered together as evidence. A simple solution to this problem is to average across all vowel phonemes to produce a LTFD. The only drawback to this lies in the high level of generalization that is entailed when all vowels are averaged, meaning that idiosyncrasies in individual phonemes may be overlooked. It appears that both LTFD and MFCC analysis can provide insights into the vocal tract; however, under an LR framework only one of these vocal tract parameters (LTFD or MFCC) would be combined with other pieces of speech evidence into an overall LR (owing to the strong correlations between LTFD and MFCC; French et al., 2012). For this reason, unless a single phoneme can yield more promising LR results for different populations, these results suggest that LTFD should be considered over individual vowel analysis under the LR framework.

# Chapter 6: Long-Term Fundamental Frequency

## 6.1 Introduction

Long-term fundamental frequency (F0) is a commonly-used feature in forensic speaker comparisons. Clark and Yallop (2001, p. 332) define fundamental frequency as “the number of times per second that the vocal folds complete a cycle of vibration.” Long-term fundamental frequency is the measure of fundamental frequency over longer segments of speech, instead of smaller intervals (e.g. a phoneme, a word). Clark and Yallop further explain that F0 is “controlled by the muscular forces determining vocal fold settings and tensions in the larynx, and by the aerodynamic forces of the respiratory system which drive the larynx and provide the source of energy for phonation itself” (Clark and Yallop, 2001, p. 333). Speakers are known to differ from one another in the distribution of spectral energy (of F0) within their speech, due largely to anatomical reasons and the way in which individuals manage their phonatory/vocal tract settings (Clark and Yallop, 2001). For this reason, F0 is commonly analyzed in forensic speaker comparisons, with the aim of identifying those speaker-specific differences found in vocal fold vibrations and phonatory and other vocal settings. The survey completed by expert forensic phoneticians discussed in Chapter 3 reports that all experts considered F0 in casework. Alongside voice quality, F0 was also claimed to be the most useful speaker discriminant by experts. The most commonly-measured aspects of F0 were mean and standard deviation (§ 3.9.1.2).

Despite the popularity of F0, the parameter is not immune to exogenous factors such as emotion, disguise, alcohol, drugs, telephone transmission, and recording codecs (Braun, 1995; Gold, 2009; Künzel, 2001; Papp, 2008). It is already highly variable within speakers and the presence of these factors makes it even more so. Regardless of this, however, the experts' expectations remain that it is useful in discriminating between speakers. An example of F0 playing a key role in a forensic case is outlined in Nolan (1983, p. 124). The expectation of F0 being a good speaker discriminant may stem from the view that it is "to some extent anatomically determined" (Hudson et al., 2007, p. 1809). On a positive note, F0 has been shown to be rather robust to background noise and is not greatly affected by telephone transmission (Braun, 1995).

In order to evaluate experts' expectations regarding the discriminant power of F0 (expressed in terms of mean and standard deviation), empirical testing is required on large homogeneous groups of speakers. There have been a number of studies examining F0 in English (Hudson et al., 2007; Graddol, 1986; Loakes, 2006). However, only Hudson et al. (2007) provides statistics for a group of English speakers. There is also only one study on English that reports on within-speaker variability. However, this is for a set of only eight speakers of Australian English (Loakes, 2006). For this reason, the current chapter examines inter- and intra-speaker variation in F0, and considers the discriminant potential of F0.

## 6.2 Literature Review

Fundamental frequency has previously received a large amount of attention in forensic phonetic research. Kinoshita et al. (2009, p. 92) suggest that the popularity of F0 “stems from promising results in early speaker recognition research” (such as the work by Atal (1972)), as well as F0 fitting three of Nolan’s standards for good forensic speaker comparison features: its robustness, measurability, and availability (Nolan, 1983). This section brings together relevant F0 literature, and divides it into two parts. The first part considers F0 in general as a population statistic and a speaker discriminant in forensic phonetics, and the second examines exogenous factors that can affect F0 and its measurement.

### 6.2.1 F0 as a Speaker Discriminant in Forensic Phonetics

Research on long-term F0 has resulted in a number of published statistics that are often cited as reference data, especially in relation to forensic speaker comparison casework and research. Fundamental frequency statistics for male and female speakers in both read and spontaneous speech, and across multiple languages, are provided in Traunmüller and Eriksson (1995). Many new studies have been conducted since Traunmüller and Eriksson (1995). The majority of those that have had forensic motivations have specifically analyzed F0 in spontaneous speech, and among larger and/or more homogeneous groups of speakers.

Rose (2003) reports long-term F0 measurements for non-contemporaneous read speech (recordings separated by approximately one year), produced by six male speakers of Australian English. The six speakers had a mean F0 in the first recording of 113.6Hz (range: 101.9-124.8Hz) and a mean standard deviation of 21.7Hz (range: 15.24-30.5Hz). The second recording had a mean F0 of 114.5 Hz (range: 101.4-127.6Hz) and a mean standard deviation of 17.4Hz (range: 14.3-19.2Hz). Loakes (2006) also presents F0 values for male speakers of Australian English. Measurements of F0 were taken at the midpoints of vowels in eight-minute recordings. The eight speakers had a mean F0 of 105.2Hz, a median F0 of 103.1Hz, a mode F0 of 107Hz, and a mean standard deviation of 16.4Hz. The F0 values for spontaneous speech in Loakes (2006) were lower than those reported by Rose (2003). However, read speech tends to elicit higher F0 values than spontaneous speech (Loakes, 2006).

Lindh (2006) reports long-term F0 values for 109 young male Swedish speakers taken from short samples of spontaneous speech. The male speakers, aged 20-30, had a mean F0 of 120.8Hz, a median F0 of 115.8Hz, and an average alternative baseline F0 of 86.3Hz. The alternative baseline is the value (in Hz) that falls 7.64% below the mean F0 (approximately 1.43 standard deviations; see Lindh (2006) for more on alternative baseline). These F0 values are higher than those found for Australian English. Rose (2002) suggests that F0 values may be language-specific. The findings presented by Lindh (2006) also indicate that collecting F0 statistics is necessary for different languages and perhaps even different dialects/accents in order to understand the significance of specific F0 measurements in forensic casework.

The study most relevant to the research reported in the current chapter is that by Hudson et al. (2007), which investigates long-term F0 in the speech of 100 male speakers of British English drawn from the DyViS database. The authors use Task 1 of DyViS, where individuals are taking part in a simulated police interview. Three to five minutes of spontaneous speech per speaker were analyzed after all background noises were removed. A Praat script was used to extract mean, median, and mode F0 for each speaker. The aim of the research was to gain an understanding of the distribution of F0 in a large homogeneous group of English speakers primarily for forensic casework. Hudson et al. (2007) report a group mean F0, median F0, and mode F0 of 102.2, 106, and 105 Hz respectively. The study provides the population results in a format that is useful for interpreting data with respect to between-speaker variation. However, it does not offer any analysis of within-speaker variation. Under a numerical LR framework it is necessary to provide an analysis of intra-speaker variation in addition to the more commonly-studied inter-speaker variation. As such, Hudson et al. (2007) cannot attach a numerical value to the level of discrimination that can be achieved using only F0 as a discriminant.

The studies detailed above have reported average F0 values for groups of speakers, while ignoring individual speaker variation (aside from Loakes, 2006 and Rose, 2003, but at a very limited level). Kinoshita (2005) was the first to investigate intra- and inter-speaker variation in F0 in conjunction with the discriminant power of F0 on a large scale. Kinoshita provides long-term F0 statistics derived from non-contemporaneous samples of spontaneous speech for 90 male speakers of Japanese, reporting a mean F0 of 135.7Hz and a standard deviation of 26.4Hz. Likelihood ratios were calculated using Lindley's (1977) formula and synthetic (i.e. invented) criminal and suspect F0s (both mean and SD) were created. The 90 male speakers acted as the reference population.  The results presented had a small range of LR estimates, and were rather close to unity (i.e. not supporting a preference for one hypothesis or the other). Kinoshita (2005) therefore suggested that long-term F0 is not a very strong speaker discriminant and that it contributes very low strength of evidence.

### 6.2.2 Effects of Exogenous Factors on F0

Intra-speaker variation is caused by numerous factors, such as intoxication, emotion, vocal effort, disguise, and recording/transmission technology, to varying extents (Braun, 1995; Gold, 2009; Junqua, 1996; Künzel, 2001; Liénard and Di Benedetto, 1999; Papp, 2009; Zetterholm, 2006). F0 has been shown not to be as robust as previously believed, and it is sometimes the case in forensic cases that F0 measurements are not robust to the effects of these exogenous factors (Peter French, p.c.).

Braun (1995) draws attention to a number of factors known to affect F0 and which may have some relevance to forensic situations. An exhaustive list of known effects on F0 is presented with relation to technical, physiological, and psychological factors. She gives, as examples of technical factors, tape speed, electronic voice changers/disguise (Hollien and Michel, 1968; Künzel, 1987), and sample size (French, 1990; Horii, 1975; Mead, 1974; Steffan-Battog et al., 1970). Physiological factors affecting F0 include speaker race (Hudson and Hollbrook, 1981), age, a history of smoking (Braun, 1994; Gilbert and Weismer, 1974; Murphy and Doyle, 1987; Sorensen and Horii, 1982), alcohol consumption (Klingholz et, al. 1988; Künzel et al., 1992; Pisoni and Martin, 1989; Sobell et al., 1982), testosterone drugs and anabolic steroids (Bauer, 1963; Berendes, 1962; Damasté, 1964; 1967), removal of cysts/nodules/polyps (Bouchayer and Cornut, 1992), surgical stripping of the vocal folds after edema/tonsillectomy/thyroidectomy (Ardnt, 1963; Fritzell et al., 1982; Keilmann and Hülse, 1992), shortening of vocal folds (Oats and Dacakis, 1983), and lingual block (e.g. use of anesthesia; Hardcastle, 1975). Finally, Braun (1995) identifies a number of psychological factors known to affect F0, which are emotions (sorrow, anger, fear; Williams and Stevens, 1972), stress (Hecker et al., 1968, Scherer, 1977), vocal fatigue (Novak et al., 1991), depression (Darby and Hollien, 1977; Hollien 1980; Scherer et al., 1976), schizophrenia (Hollien, 1980; Saxman and Burk, 1968), time of day (Garrett and Healey, 1987), and background noise level (Dieroff and Siegert, 1966; Lombard, 1911; Schultz-Coulon, 1975; Schultz-Coulon and Fues, 1976). These extensive lists detailed by Braun (1995) pose problems that FSC experts are confronted with when analyzing F0. Many other studies have been carried out since Braun (1995) to examine the effects of external factors on F0. Those studies that are most relevant to casework are detailed below.

Liénard and Di Benedetto (1999) examined the effects of vocal effort on F0. They looked at 12 French vowels spoken in isolation by ten speakers (five males and five females). Vowels were repeated multiple times to an experimenter who stood at varying distances in the room from the speaker (close, normal, and far). The distance at which the experimenter stood relative to the speaker was intended to induce change in the vocal effort that the speaker assumed would be required for the experimenter to hear the speaker clearly. Liénard and Di Benedetto (1999) found F0 to increase by around 5Hz.

Voice disguise is another common cause of variation in F0. Künzel (2000) reports that nearly 25% of cases in Germany involve voice disguise, and he specifically investigated the effect of such disguise on F0. He analyzed read speech from 100 speakers (50 males and 50 females) where they were asked to adopt different voice disguises (high, low, and denasalized). Künzel showed that speakers were effectively and consistently able to disguise their voices using F0 modulation, some to extreme levels. Most importantly, he notes that individuals were observed applying different phonatory strategies, which caused difficulty in associating the change in F0 with a particular change the speaker has made in his/her phonatory setting. Zetterholm (2006) has also examined disguise and imitation of voices in relation to changes in F0 using the speech of a single individual (a professional impersonator), impersonating 12 different popular Swedish TV personalities. She found that the impersonator was able to vary his modal F0 voice from its normal average of 118Hz, to anywhere between 97Hz and 225Hz.

The effects of recording and transmission technology on speech has received more attention of late, perhaps due in part to the advent of new and emerging technologies. This opens new avenues for potential technical effects on F0 and on speech in general. Gold (2009) considers one such area, by investigating the effects of video and voice recorders in cellular phones. Three different cellular phones were used in the experiment. All phones encoded the speech signal using an AMR or mpeg4 codec for voice and video recorders respectively. A change in the speakers' F0 was found to result in differences of between one and five percent in mean F0, and the SD of F0 changing from 9 to 63.6% for a single cellular phone. Overall, voice recorders (AMR codec) were found to make bigger changes in F0 than video recorders did. It has previously been shown that the GSM AMR codec (a speech-encoding codec similar to the AMR) has the tendency to change voiced frames into unvoiced frames, and vice versa, which in turn affects F0 measurements (Guillemin and Watson, 2008, p. 216). This could be the case for the AMR codec found in cellular phones, and any recording devices in general that incorporate a similar-functioning codec to the GSM AMR.

Variation in F0 can be caused by numerous exogenous factors, many of which (those relevant to casework, at any rate) have been detailed above. The variability of F0 due to external factors introduces many difficulties when comparing recordings that have been affected by different combinations of exogenous factors. For this reason it is important to consider the effects of within- and between-speaker variation when investigating the discriminant potential of F0.

## 6.3 Population Statistics for Fundamental Frequency

The following section presents population statistics for F0 in a large, linguistically homogeneous group of 100 male speakers. These data serve as the first of their kind in providing detailed information on intra-speaker variation in F0 among individuals who speak Southern Standard British English (SSBE). Additionally, population data are provided for between-speaker variability in F0, where in the discussion it will be made apparent that the variability is similar to that found in Hudson et al. (2007).

### 6.3.1 Methodology

The current study uses the recordings from Task 2 of the DyViS database. Each recording for all 100 speakers was used in its entirety. However, after the editing of the files, the recordings were between 2:25 minutes and 11:17 minutes in length, with an average of 6:21 minutes per file. Using Praat (version 5.1.35), multiple passes were made through the recordings to ensure that there was only speech remaining. The first phase consisted of the removal of all the portions where the interlocutor was speaking, and any silent pauses with the talkers' speech. The second phase removed all intrusive noises in the recordings, including background noises, laughter, coughs, and sneezes. A final listening phase was used to ensure that everything but the speech of the speaker had been properly removed from the recordings. This final phase saw only minor edits that typically amounted to the removal of less than one second of net speech per speaker.

As the calculation of LRs requires multiple tokens per speaker, it was necessary for all recordings to be divided into segments in order to establish within-speaker variability. The amount of net speech necessary for a given F0 token has not been previously tested. For this study I chose to segment the speech into 10-second intervals (see § 6.4.3 for effects of package length). Each file was then annotated using a Praat text grid. The tier represented the package length (in seconds) according to which the speech signal was subsequently segmented. Starting from the beginning of each recording, intervals were marked out in the text grids until the end of each recording. If the final segment did not meet the interval length requirement, it was not included in analysis. Figure 6.1 depicts an example of the text grid annotations used for all recordings.

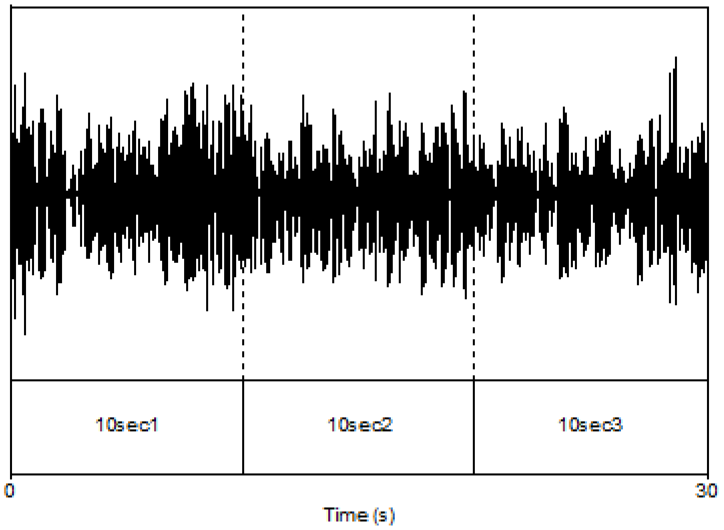


**Figure 6.1:** Example of a text grid annotation

A Praat script (pitch_mean_results.txt; created by Henning Reetz, 2009) was used to extract mean F0 and standard deviation values for each interval. The Praat script was set to a frequency range of 50 – 300 Hz (following Hudson et al. 2007). After reviewing the F0 Praat picture distributions (for octave jumps and unwanted pitch artefacts), 64 speakers were found to have reliable F0. The remaining 36 speakers' F0 values contained obvious errors (e.g. octave jumps), and were therefore re-run using tailored ranges. The tailored ranges were chosen through trial and error, where the range with the least amount of errors was chosen as the best possible frequency range. These are detailed in Table 6.1.

**Table 6.1:** Tailored Frequency Ranges for Selected Speakers

| Frequency Range (Hz) | Speaker |
|---|---|
| 50-150 | 024, 026, 050, 062 |
| 50-160 | 074, 081, 099 |
| 50-200 | 052, 056 |
| 55-160 | 007, 009, 025, 029, 036, 040, 073, 096 |
| 55-200 | 045, 049, 055, 065, 075, 078, 082 |
| 75-200 | 003, 015, 021, 035, 041, 058, 059, 066, 083, 092 |
| 75-160 | 006, 014 |

After the Praat script was re-run using tailored frequency ranges for these speakers, all F0 means and standard deviations were imported into Microsoft Excel for further analysis. There were a total of 7,447 intervals for all 100 speakers. On average each speaker had 74 intervals.

### 6.3.2 Results

The distributions of F0 mean and standard deviation for individuals are presented in Figures 6.2 and 6.3. The y-axis represents the number of speakers that fall within a given range and the x-axis depicts F0 in Hertz (Hz).

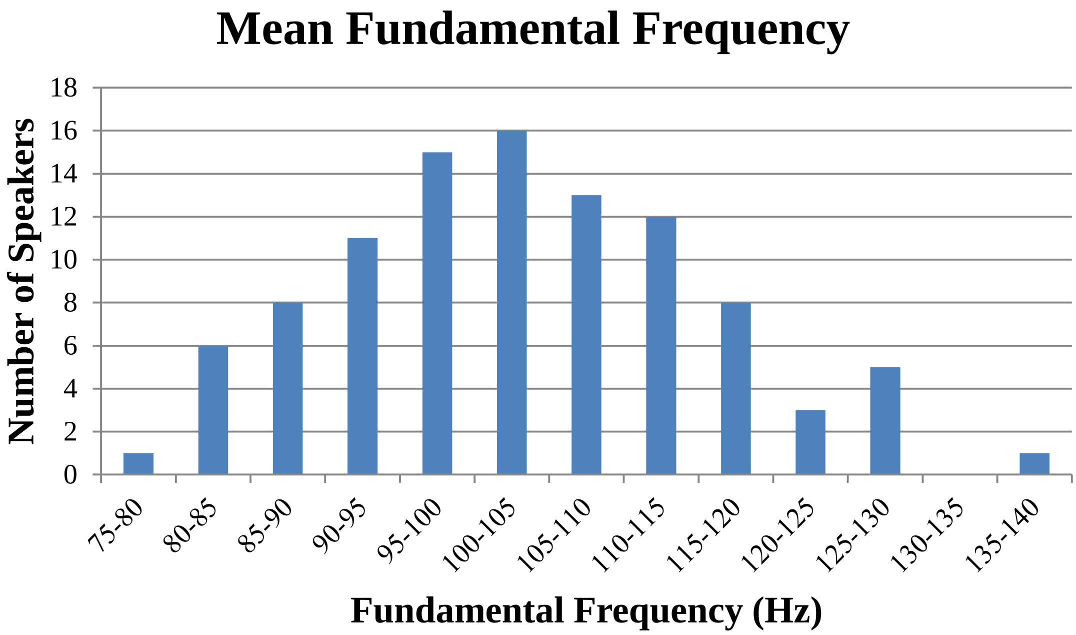


**Figure 6.2:** Distribution of mean fundamental frequency

There is a normal distribution in the DyViS corpus for mean F0[37], as illustrated in Figure 6.2. The mean F0 for the population is 103.6Hz, with an overall mean F0 range of 79.9Hz-136Hz. The standard deviation of the means is 12.77Hz. There are no suspected outliers (as defined in Chapter 4) in the mean F0 data.

[37] Technically, this is the mean of the means of the means (i.e. the mean across speakers of the means across tokens of each speaker of the means of all the raw F0 values of each token).

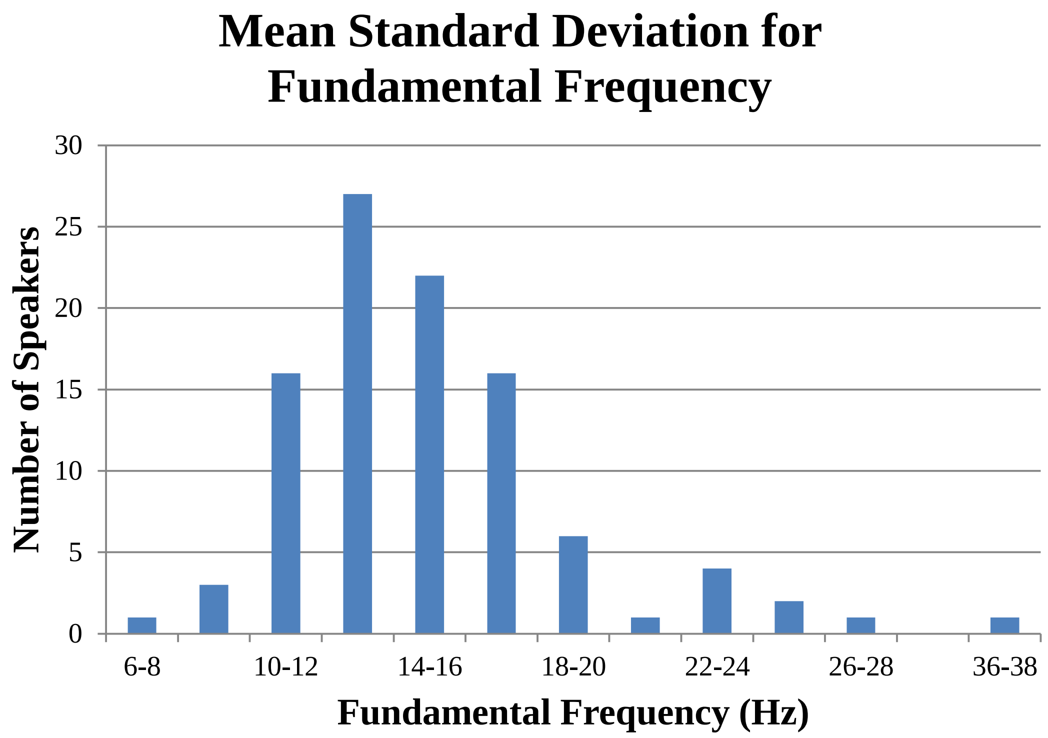


**Figure 6.3:** Distribution of standard deviation in fundamental frequency

There is a roughly normal distribution for the SDs of F0[38] in Figure 6.3, with a slight positive skewing due to a number of outliers. There are seven suspected higher outliers (22.62Hz (for two speakers), 23.25Hz, 23.29Hz, 24.82Hz, 25.59Hz, and 27.62Hz) and one extreme outlier at 37.32Hz. Including the outliers, the mean SD for the population is 15.1Hz, with a range of 7.4Hz-37.3Hz. If those outliers are removed, the distribution becomes more normal and the mean SD is then 14.17Hz.

The cumulative distribution graphs of mean F0s and F0 SDs in Figures 6.4 and 6.5 (respectively) show the percentiles at which a given F0 mean or SD falls in relation to the population. The y-axis represents the cumulative

[38] Technically, this is the mean SD of the mean SD (i.e. the mean SD across speakers of the SDs across tokens of each speaker of the means of all the F0 raw values of each token).

proportion of the population, and the x-axis presents fundamental frequency (Hz).

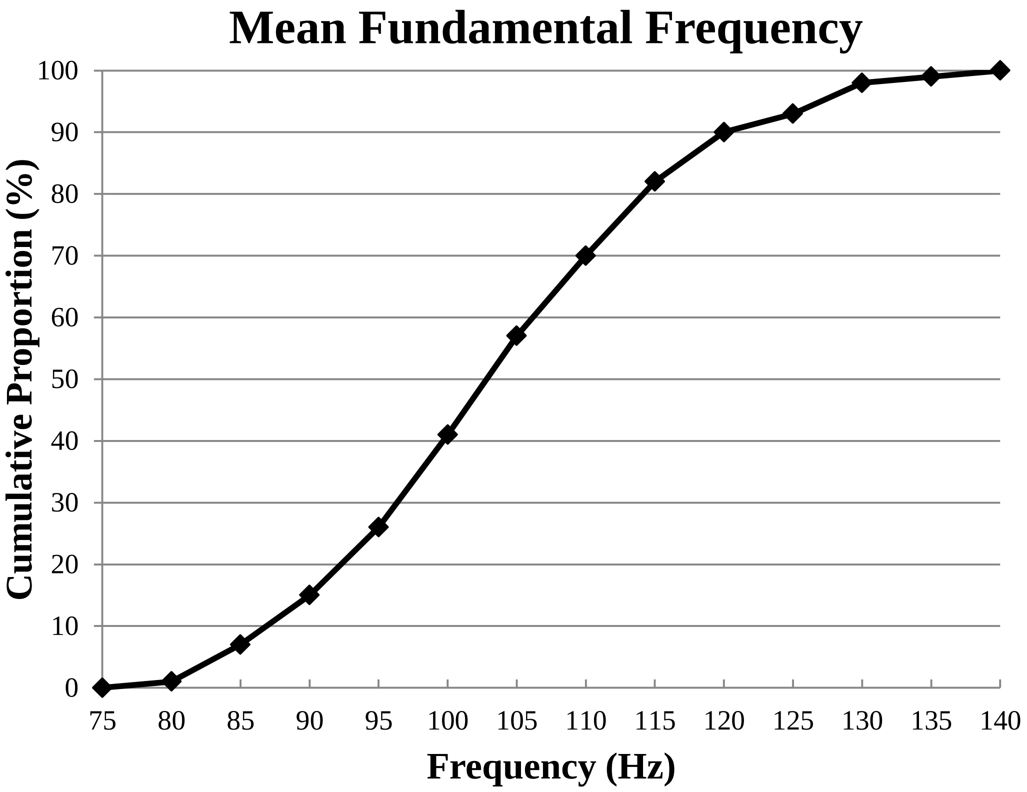


**Figure 6.4:** Cumulative percentages for mean fundamental frequency

The curve in Figure 6.4 is characterized by a steep central section but has gentle gradients at both ends. The data show that the lowest 20% of the speakers have a mean F0 below 93Hz, while the highest 20% have an F0 above 115Hz. This leaves only a narrow band of 22Hz in which the remaining 60% of speakers are found. This is indicated by the steepest portion of the trajectory.

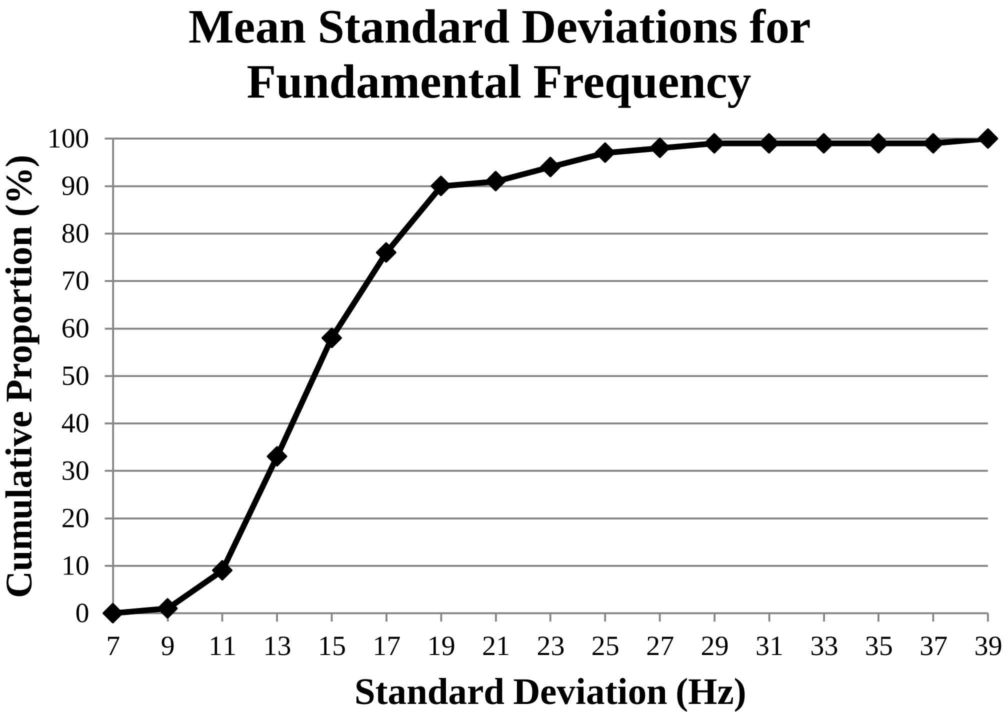


**Figure 6.5:** Cumulative percentages for standard deviation in fundamental frequency

The first half of the curve in Figure 6.5 is characterized by a steep trajectory, while the second half has a fairly long gradient trajectory at the upper end. The lengthy gradient to the end of the trajectory is due to the suspected outliers and extreme outliers confirmed above in the current section. Observing the spread of the SDs, the lowest 20% of speakers have F0 SDs below 12Hz, and the highest 20% have F0 SDs above 17.5Hz. This leaves a remarkably narrow band of 5.5Hz in which the majority of speakers fall (60%). Overall, the F0 data have a variance ratio of 0.7152, which indicates that there is more variation occurring within speakers than between them.

### 6.3.3 Discussion

The results for the present study are very similar to those reported by Hudson et al. (2007), which used Task 1 of DyViS. The difference in the two group means is only 2.4Hz. Results from the present study are also similar to those presented in Loakes (2006), who gave a mean F0 of 105.2Hz for the spontaneous speech of Australian English-speaking males. For non-English languages the results presented in § 6.3 are somewhat dissimilar in that both Swedish (Lindh, 2006) and German (Künzel, 1989) report higher mean F0s at 115.8Hz and 120.8Hz, respectively.

The lower mean F0 values found for SSBE compared to those found for other languages could in fact be a result of F0 being language-specific, as suggested by Rose (2002). However, it could potentially be caused by the fact that numerous speakers in the DyViS database have creaky voice qualities (Hudson et al., 2007), and the creaky voice qualities of speakers were included in the current study. This is because the F0 of speakers who use creaky phonations a lot tend to result in bimodal distributions, with the first peak representing the creaky voice quality and the second peak representing modal phonation. In order to calculate a mean F0 for a speaker the two phonation types are averaged, which thus results in a lower mean F0 (Hudson et al., 2007). As such, it is most likely the case that, as pointed out in Hudson et al. (2007), speaker's modes did not correspond to their means. For this reason it would be ideal to find a more accurate way of representing the mean F0 of creaky-voiced individuals. Overall, the results also suggest that mean F0 and SD are perhaps not the best measures for those speakers with intermittently-present creaky voice.

## 6.4 Likelihood Ratios

The discriminant results for F0 are presented in the following section. The first part provides LR results for F0, and the second part considers the effects of package length on LR results for F0.

### 6.4.1 Methodology

As seen in § 4.6.1, likelihood ratios were calculated using a MatLab implementation of Aitken and Lucy's (2004) Multivariate Kernel-Density (MVKD) formula (Morrison, 2007). An intrinsic methodology was used, whereby the test and the reference speakers came from the same population of 100 speakers. Speakers 1-50 were used as the test speakers, while speakers 51-100 served as the reference speakers. Mean F0 and SD parameters were both used for each token spoken by a given individual in the calculation of the LRs. Performance of the system was assessed in terms of both the magnitude of LRs (Champod and Evett, 2000) and system validity (Cllr and EER).

### 6.4.2 Results for F0

The results for the calculation of LRs for F0 are summarized in Table 6.2. The second row contains the results from same-speaker (SS) comparisons and the third row contains the different-speaker (DS) comparison results. The second column indicates the percentage of comparisons in which speakers were correctly identified, whereby a log likelihood ratio (LLR) above zero was correct for a SS comparison and an LLR of less than zero was a correct judgment for DS comparisons. The mean LLR is found in the third column, followed by the minimum and maximum LLRs. The final two columns present the performance of the system in terms of EER and Cllr, respectively.

**Table 6.2:** Summary of LR-based discrimination for F0 (100 speakers)

| Comparison | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **10 sec SS** | 92.0 | 0.958 | -3.404 | 1.936 | 0.0849 | 0.4547 |
| **10 sec DS** | 89.9 | -24.204 | -269.159 | 1.906 | | |

Table 6.2 shows that SS comparisons slightly outperform DS comparisons in the percentage of correct judgments. The mean LLR for DS offers very strong evidence to support the defense hypothesis ($H_{d;}$ Champod and Evett, 2000), while the mean LLR for SS only offers limited evidence to support the prosecution hypothesis ($H_p$). Even the Max LLR for SS does not reach a strength of evidence of 2 (instead, only moderate evidence to support $H_p$ is indicated by the value of 1.936). The EER for the system is higher than that found for a combined LTFD system in Chapter 5 (0.0414; see Table 5.4), but is significantly better than that found for AR in Chapter 4 (0.334; see Table 4.7). The Cllr for F0 as a system is generally better than the Cllrs achieved in Chapters 4 and 5 for AR and LTFD. A Cllr closer to zero would nonetheless be desirable.

The Tippett plot in Figure 6.6 offers a visual measure of the performance of F0 as a discriminant feature.

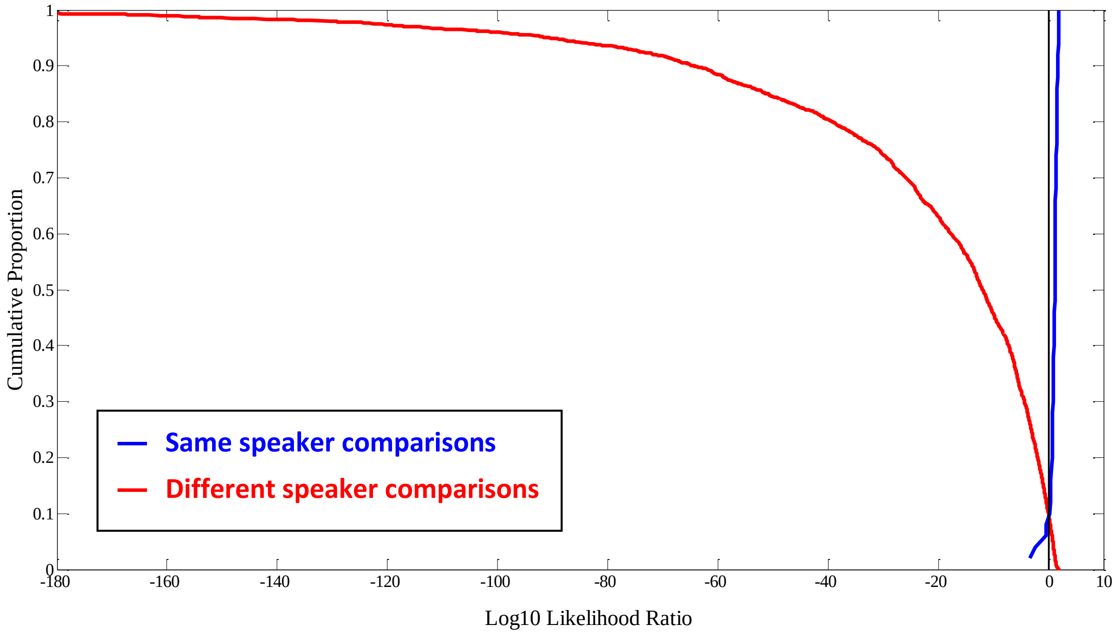


**Figure 6.6:** Tippett plot of fundamental frequency

Figure 6.6 shows that there is a narrow range in LLR for SS and that most LLRs for SS are relatively similar. The DS comparisons have a wider spread of LLRs. It is also clear that DS comparisons can achieve very large LLRs, which offers a high strength of evidence.

### 6.4.3 Results of Package Length

In order to establish variability within a speaker under an LR framework, multiple tokens of a speech parameter of an individual are needed for analysis. This involves dividing the recording into multiple sections (or tokens). The most efficacious token length (or referred to here as package length) has not been previously established. Therefore, a package length of 10 seconds was chosen for the study as it was found to yield the lowest EER. However, it is possible to vary the size of the package length (similar to that seen in Chapter 5). The effects of package length variation can be seen in Table

6.3 for F0. The table has the same formatting as that of Table 6.2, although Table 6.3 is expanded to include additional rows containing the various package lengths (5, 10, 15, and 20 seconds).

**Table 6.3:** F0 package length variability for LR results

| Comparison | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **5 sec SS** | 88.0 | 0.868 | -5.176 | 2.030 | 0.1010 | 0.5634 |
| **5 sec DS** | 91.1 | -29.879 | -308.588 | 2.031 | | |
| **10 sec SS** | 92.0 | 0.958 | -3.404 | 1.936 | 0.0849 | 0.4547 |
| **10 sec DS** | 89.9 | -24.204 | -269.159 | 1.906 | | |
| **15 sec SS** | 92.0 | 0.970 | -2.526 | 1.999 | 0.1016 | 0.4407 |
| **15 sec DS** | 89.0 | -20.964 | -233.785 | 1.809 | | |
| **20 sec SS** | 92.0 | 0.960 | -2.536 | 1.880 | 0.0967 | 0.4383 |
| **20 sec DS** | 88.7 | -18.620 | -206.961 | 1.717 | | |

The results in Table 6.3 suggest that package length affects the discriminant performance of F0. However, the increase of package length does not appear to be linearly correlated with the overall system performance in terms of EER. For Cllr, there does appear to be a direct relationship between the increase in package length and the improvement in Cllr.

Effects of package length can also be considered from inspection of the values shown in Table 6.4, which displays the results for mean F0s, F0 range, standard deviations (SD), and range of SDs across the four different package lengths.

**Table 6.4:** Fundamental frequency across different package lengths

| Package Length | Mean of Means (Hz) | Range of Means (Hz) | Mean of SDs (Hz) | Range of SDs (Hz) |
|---|---|---|---|---|
| 5 sec | 103.2 | 79.7-136.1 | 14.3 | 6.9-36.2 |
| 10 sec | 103.6 | 79.9-136.0 | 15.1 | 7.4-37.3 |
| 15 sec | 103.3 | 79.8-136.2 | 15.4 | 7.6-37.6 |
| 20 sec | 103.2 | 79.9-136.1 | 15.5 | 7.8-37.9 |

Table 6.4 shows that there is relatively little difference between F0 results as a function of the different package lengths. The biggest difference found in the results is in the mean of SDs for 5 seconds (14.3Hz), compared to the mean of SDs for the larger package lengths (15.1-15.5Hz). The 5-second package length was also found to have the biggest difference in Cllr (Table 6.3) with the longer package lengths. On the basis of these results, it could be argued that choosing a package length of 10 seconds or above will give an accurate representation of the data.

It is important to note that like the package length results found for LTFD in Chapter 5, the results presented in this section relate specifically to the present recordings, in which the total length of material per speaker was around six minutes. It could again be the case that package length affects longer or shorter speech samples differently. However, this analysis serves as a starting point for further investigation into this issue.

## 6.5 Discussion

The results presented in the present study provide a starting point for further investigation into the discriminant value of F0. However, the study was limited by the highly controlled nature of the recordings, which were relatively free from the influences of the exogenous factors that are known to affect F0 values, as detailed in § 6.2.2. More studies which incorporate those factors are needed.

The results of the present study were produced using only mean and standard deviation as discriminant parameters of F0. This choice was dictated by opinion given in the survey completed by expert forensic phoneticians. The survey also reported that it is not uncommon for experts to use other measures of F0 in their casework. Kinoshita et al. (2009) showed promising results when using parameters that described the distribution of F0 more precisely (skew, kurtosis, modal F0, and modal density). This points to a need to reassess the measures commonly used in relation to F0. Different measures of F0 could lead to more detailed descriptions of F0 distributions that also achieve a higher strength of evidence.

## 6.6 Conclusion

The results presented in this chapter suggest that F0 is a moderately good speaker discriminant overall, and has promise for demonstrating that two voices have come from the same speaker (rather than different speakers) in the same recording (achieving an EER of 0.0849). However, it is not known how well F0 can discriminate between individuals when same-speaker evidence comes from different recordings. Previous literature would suggest that its discriminant potential will decrease when same-speaker evidence from different recordings is introduced (see § 6.2.2). F0 as a speaker discriminant showed more variation occurring within speakers than between speakers. However, F0 does show there to be more variation present between speakers than do AR (Chapter 4) and individual LTFDs (Chapter 5).

It is difficult to ascertain whether experts responding to the survey were correct in identifying F0 as a good speaker discriminant. Results suggest that they are correct in that F0 does well discriminating same speakers that come from the same recording, but it is uncertain whether that result will hold true when same-speaker comparisons involve different recording sessions or introduce degrading factors (e.g. disguise, intoxication, background noise). The good recording conditions and high audio quality used for the current study are not reflective of those found in real casework.

Owing to the many exogenous factors detailed in § 6.2.2, the mere comparison of mean F0s and SDs is unlikely to advance the methods used for the speaker comparison task dramatically on its own. However, as always, exceptions are to be made for those individuals who lie towards the margins of the distribution curve or who can be classed as outliers, and the case remains for using F0 in conjunction with other speech parameters.

# Chapter 7: Click Rate

## 7.1 Introduction

The survey results presented in Chapter 3 indicated that experts examine many non-linguistic features as part of their analysis in FSCs. Those non-linguistic features can include patterns of audible breathing, laughter, throat-clearing, *tongue clicking*, and both filled and silent hesitation phenomena. In respect of tongue clicking, Chapter 3 shows that 57% of the practitioners questioned examined recordings for the presence of velaric ingressive stops (i.e. clicks), and 18% considered them to be a highly discriminant feature.

Research into the discriminant ability of parameters in forensic speech science has focused primarily on vowels, and to some extent consonants and fundamental frequency (Gold and Hughes, 2013). However, there remains a gap in the literature pertaining to the discriminant ability of non-linguistic parameters (e.g. clicks).

This chapter investigates the speaker discriminant power of clicks, which are defined here as a linguistic parameter rather than a non-linguistic parameter (reported in Chapter 3 as non-linguistic). This is because the clicks analyzed in this chapter are used by speakers in a discursive manner that can be classified as conveying linguistic meaning (i.e. they are used here as a discourse marker in conversation). The first part of this chapter investigates the discriminant power of clicks by analyzing population statistics for click rate, and the second portion analyzes the robustness of clicks in relation to accommodation effects. The final limitation section in this chapter is devoted entirely to discussion of calculating likelihood ratios for clicks, and the difficulties in attributing a numerical strength of evidence to measure discrete data.

## 7.2 Literature Review

Ladefoged (2006, p. 292) defines a click as “a stop made with an ingressive velaric airstream, such as Zulu [ǁ].” Laver (1994, p. 174) explains further that, “a major ingredient in the production of the airstream [for clicks] is a complete closure made by the back of the tongue against the velum. A second closure is also made, further forward in the mouth, either by the tip, blade or front of the tongue, or by the lips.” For a lingual click, there is a closure made by the back of the tongue coming into contact with the soft palate, and the front portion of the tongue is then drawn downwards. This process increases the volume of the space occupied by the air trapped in between the two closures “rarefying the intra-oral air-pressure. When the more forward of the two closures is released, the outside air at atmospheric pressure flows in to fill the partial vacuum” (Laver 1994, p. 174). It is at this point that a click is realized. Figure 7.1 below illustrates the actions of the vocal organs involved in the production of a click sound.

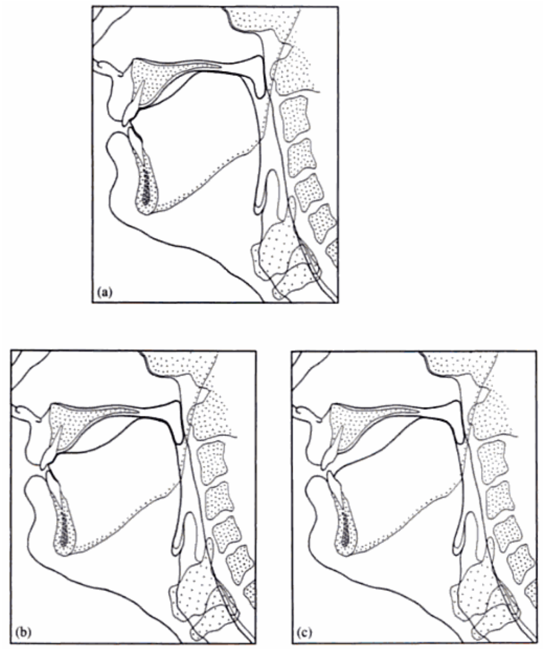
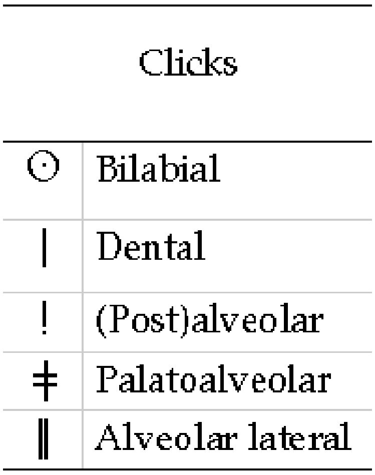


**Figure 7.1:**) "The action of the vocal organs in producing a velaric ingressive voiceless dental click [k͡ǀ]: (a) first stage, velic and anterior closure; (b) second stage, expansion of the enclosed oral space; (c) third stage, release of the anterior closure." (Laver, 1994, p. 176)

Figure 7.1 provides an illustration of the process involved for the vocal organs in the production of a dental click. This is just one of six possible places of articulation for clicks as recognized by the International Phonetic Association (IPA). The five different click types are provided in the figure below, which is an extract from the IPA chart.



| Clicks | |
|---|---|
| ʘ | Bilabial |
| ǀ | Dental |
| ! | (Post)alveolar |
| ǂ | Palatoalveolar |
| ǁ | Alveolar lateral |

**Figure 7.2:** IPA Chart - Clicks Excerpt

The six different clicks presented in Figure 7.2 above are most commonly recognized for their existence in a number of African languages (Ladefoged 2006, p. 139) and extensive research has been carried out to document clicks in those languages (e.g. Greenberg, 1950; Herbert, 1990; Jessen and Roux, 2002). In African click languages, such as in Xhosa, Zulu, Sandawe, Hadzapi, Bushman, Nama, !Xóõ, and !Xũ clicks are used phonemically (Laver, 1994, p. 174). Clicks are also found in English, but unlike those in African languages they are not used phonemically. According to literature on clicks found in English, they have typically been described as functioning on only a paralinguistic level to denote the attitudes, intentions (e.g. encouraging a horse to move), and emotional states of a speaker. Previous research suggests that certain clicks are used to convey such things as annoyance (Abercrombie, 1967, p. 31; Ball, 1989, p. 10), sympathy (Gimson, 1970, p. 34), and disapproval (Crystal, 1987, p. 126). There is also evidence to suggest that the phonetic properties of clicks can vary depending on their functions in English (Gimson, 1970, p. 34).

Wright (2005; 2007; 2011a; 2011b) presents an extensive amount of research focused on clicks from a non-paralinguistic point of view, specifically from a conversation analyst's view. Wright proposes three different classifications of click used in English conversation to index different meanings. The first type are clicks that occur in the onset of a new sequence, the second are clicks used in the onset of a new and disjunctive sequence, and the third type are clicks produced in the "middle of a sequence of talk, when the speaker is engaged in the activity of searching for a word" (2005, p. 176). The following are three examples from Wright (2005) of click types:

***Fragment 1: Holt.SO.88.1.2/bath/***

01: Bil: Hello:

*02: Gor: Hi Bill*

*03: Bil: Hi Gordy*

04: Gor: [ʘ] uh:m (0.4) are you going tonight

05: (.)

06: Bil: mm

07: Gor: .hhh (0.2) would you mind giving me a lif[t

08: [no that's alright

***Fragment 2: Holt.1.8/Saturday/***

01: Les: so he had a good inni:ngs did[n't he

02: Mum: [I should say so: yes

03: (0.2)

04: Mum: marvellous

05: Les: [!]. .hhh anyway we had a very good evening o:n saturday

06: (0.2)

07: Mum: Ye:s

***Fragment 3: Holt.U.88.2.2/natter/***

01: Les: .hhhh and there's the- the natte- uhm (0.2) ʘ (0.3) !

02: oh what's it called the natterjack's not so good now

Fragment 1 is an example of a click being used to start a new sequence, as noted by the bilabial click in line 4. A second type of click is used for the onset of a new disjunctive sequence in Fragment 2, which is illustrated by the alveolar click on line 5. And the final click type is found in Fragment 3, where both bilabial and alveolar clicks are being used to signify the search for a word by the speaker in line 1. The three click types above, in combination with clicks being used paralinguistically (i.e. to show emotion or affect), are used in the analysis of click productions for this chapter.

In investigating clicks in respect of their speaker-discriminating potential one can begin with the assumption that for any aspect of affect or interaction management there is no homological function-form relationship. For example, while one *can* signal annoyance, disapproval or sympathy by use of clicks, there are many other ways of signaling these states to interlocutors. Similarly, although clicks *may* be used to signal disjunction of conversational topic or the fact that one is having difficulty finding a word, other forms – semantically empty sounds or lexical expressions – can also fulfil these functions. In other words, there is an element of speaker choice in the selection of clicks over other possibilities in conveying emotive and attitudinal meaning as well as in respect of topic organization and conversational turn management. Given that this is so, one might reasonably expect there to be variability across speakers in terms of whether clicks or other forms are their preferred option. The possibility of such individual preferences provides a plausible theoretical motivation for the observation made by the forensic practitioners surveyed in Chapter 3 to the effect that clicks have high value as speaker discriminants. However, while the proposition is credible and is no doubt based on practitioners' casework experience, it has not to date been subjected to formal, empirical testing. The present chapter is an attempt to establish the speaker discriminant value of one aspect of clicking behavior, namely frequency of clicking, by such testing.

## 7.3 Data

The recordings analyzed were of 100 male speakers of SSBE aged 18-25 years from the Dynamic Variability in Speech (DyViS) corpus (Nolan et al., 2009). Two data sets were used, each of unscripted speech from a simulation of a forensically-relevant situation. One (Task 1) was a mock police interview. Each of the 100 speakers played the role of a criminal suspect and was interrogated by one of two project interviewers (Int2 and Int3) who played the role of a police officer investigating the interviewee's supposed involvement in a crime. The second set of recordings (Task 2) was of the subjects telephoning an 'accomplice' and explaining what had occurred in the police interview. The role of the accomplice in this data set was played by the same project interviewer (Int1) throughout. Although these were telephone conversations, the recordings used for analysis were made at the subjects' end of the line, i.e. they were of studio rather than telephone quality.

## 7.4 Methodology

For a feature to function as a good speaker discriminant, it must meet two criteria: (a) it must vary (ideally quite widely) across speakers; (b) it must be relatively stable within the speech production practices of individual talkers. In this section, the methods employed to test the intra- and inter-speaker variation of click rates are outlined. Task 2 recordings are used for the first portion of the click analysis of the current study. As mentioned above, each speaker conversed with a single interlocutor, Int1.

The first two minutes of each recording were ignored so as to allow for speakers to settle into the interaction. All subsequent speech from each subject, up to a maximum of five minutes net – i.e. after excluding long pauses and the speech of the interlocutor – was then extracted and divided into one-minute intervals (giving a combined total of 499 minutes of net speech). 99 of the 100 speakers produced enough speech to meet the five-minute target. One speaker (speaker 012) fell just short of this, and the analysis was therefore based on analysis of just four minutes of his speech. The extracted speech was examined auditorily during two listening sessions in Sony Sound Forge (version 10.0; analysis done auditorily) and Praat (version 5.1.35; auditory and acoustic analysis done simultaneously) for instances of clicks. Any sounds that auditorily and visually resembled clicks but were not apparently produced on a velaric ingressive airstream were excluded from the analysis. This resulted in the exclusion of 293 candidate sounds that were judged to be purely percussive[39]. At the end of this process there were a total of 454 clicks left. Each click was assigned to a functional category: either it functioned to convey affective meaning, or fulfilled one of the interactional functions identified by Wright (2007; 2011a; 2011b), i.e. initiating a new speaking turn, indicating topical disjunction, or signaling that the speaker was searching for a word. Illustrations of these interactional functions are provided in the following transcribed excerpts from the recordings (clicks are indicated by the symbol **!**, regardless of actual place of articulation):

[39] Pike (1943, p. 103) says "percussors differ from initiators in several ways: in opening and closing they move perpendicularly to the entrance of the air chamber . . . ; they produce no directional air current, but merely a disturbance that starts sound waves which are modified by certain cavity resonators; they manifest their releasing or approaching percussive timbre only at the moments of the opening and closing of some passage . . ." Typical percussives are made by the opening and closing of the lips, the tongue making closure at the alveolar ridge, the velum closing, the vocal folds making a glottal closure, and the sublaminal percussive of the 'cluck click' (Ogden 2013, p. 302). The most common percussives found in the current data set were related to the opening and closing of the lips.

***Initiating new speaking turn***

Int1: He's a tour guide now you see

→ S007: **!** Yeah, yeah, that's right and

Int1: Bear Pub

S011: Mhm

Int1: Do I know it

→ S011: **!** Um it- near Harper Passage

***Indicating topical disjunction***

Int1: Um wha- did they trace that phone call when you were in the uh grotty booth

→ S033: They asked me about it so I guess they probably have ! um but um as I wasn't, as I was telling them, I didn't go through Parkville

***Signaling word search***

Int1: And um did you give her address

S086: Uh yeah I did

Int1: Just, you know, just refresh my memory

→ S086: Yeah, sorry on Dexter Road **!** um in Dixon

Int1: Dixon this little village of Dixon

## 7.5 Results

Before addressing the central questions of inter- and intra-speaker variation, some general findings on phonetic and functional aspects of the clicks are presented.

### 7.5.1 Phonetic Properties of the Clicks

The scope of the present study did not extend to a detailed analysis of the phonetic and acoustic properties of the clicks. However, in terms of their place of articulation, approximately 95% were judged to be apical. With regard to the passive articulator, they ranged from dental through alveolar to post-alveolar. Through my own observation, dental clicks are characterized by a longer and less well-defined release phase and by a higher-frequency center of gravity and lower level of intensity than the other variants. At the other extreme, post-alveolar clicks are the highest in intensity, have a relatively short release and a greater concentration of energy at the lower frequencies. Without wishing to prejudge the outcome of further work being undertaken on these data, it appears that, at this stage at least, place of articulation proved very difficult to classify more finely, and that no individual speaker clearly stood out from the others in respect of this dimension. It is supposed that because clicks are not used phonemically by SSBE speakers, the precise place of articulation for clicks does not matter to a speaker or listener when used in sequence management. It is rather that the presence of any form of apical click can signify sequence management in conversation. Place of articulation, however, does play an important role for those clicks used as affective markers, since place of articulation for clicks has been shown to signify different emotions (Ball, 1989; Crystal, 1987; Gimson, 1970).

### 7.5.2 Functional Aspects of the Clicks

The distributions of clicks against affective function and the three interactional functions are represented in Figure 7.3.

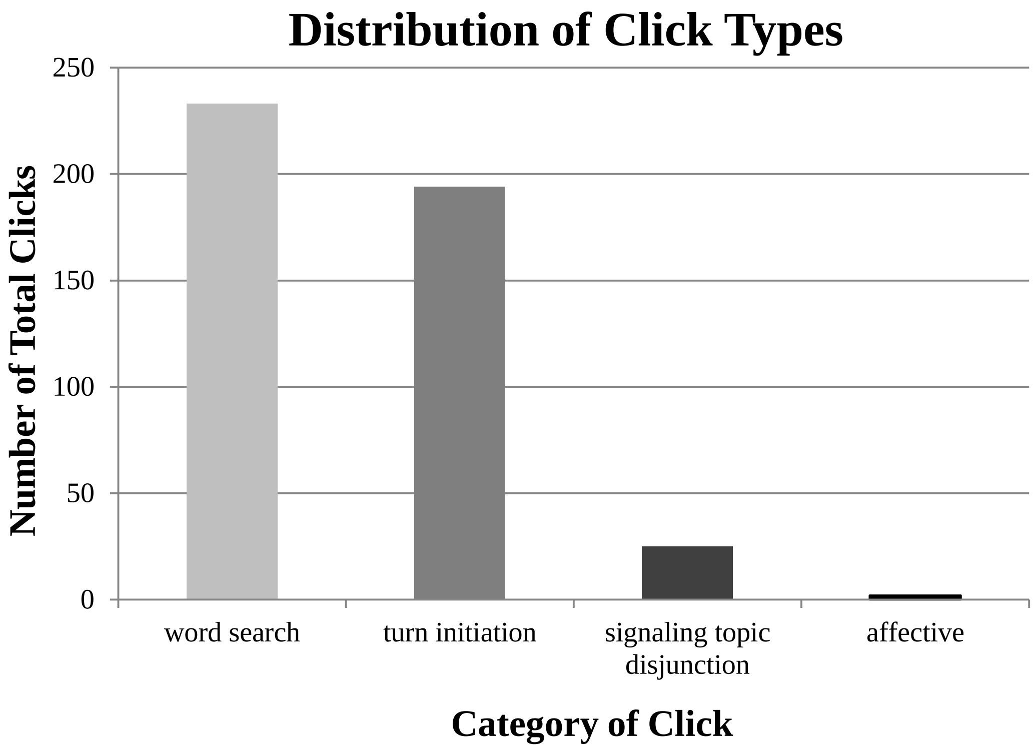


**Figure 7.3:** Distribution of click occurrences by functional category

Of the 454 clicks that occur in the combined 499 minutes of speech examined, word search accounts for just over half of all clicks (51.32%). Taken together, turn initiation and disjunction signaling clicks represent a similar proportion (48.24%) to those used to indicate word search. Affective use represents the smallest category, with only two examples (0.44%). Whilst the latter may to some extent be accounted for by the fact that the attitudinal stances that clicks are used to convey (pity, disapproval) seldom arise in the type of conversation represented in the DyViS recordings, it is nevertheless of interest that the least frequently-occurring function of clicks in these data is the one that is most frequently mentioned in the phonetic literature.

### 7.5.3 Results: Inter-Speaker Variation

The results of inter-speaker variation in click production are presented in Table 7.1, looking first at clickers versus non-clickers. The leftmost column in

Table 7.1 presents the length of time over which clicks were analyzed, while the second and third columns represent the numbers of speakers who were found to be either clickers or non-clickers. A clicker is defined as a speaker who has been found to click at least once in the given speech sample, and a non-clicker is defined as a speaker who does not click at all in the given speech sample.

**Table 7.1:** Number of clickers versus number of non-clickers over varying speech sample lengths

| **length** | **clicker** | **non-clicker** |
|---|---|---|
| *1 minute* | 39 | 61 |
| *2 minutes* | 56 | 44 |
| *3 minutes* | 67 | 33 |
| *4 minutes* | 72 | 28 |
| *5 minutes* | 75 | 25 |

As seen in Table 7.1, if one considers each sample in its entirety, the proportion of clickers to non-clickers is around 3:1 (75:25). However, this proportion could not be arrived at by examining a shorter sample, as the number of non-clickers decreases as sample length increases, owing to the fact that so many of the speakers click very infrequently. This can be seen in Figure 7.4, in which it is apparent that 74% of the DyViS population clicks five times or fewer over the five-minute period, i.e. they have a click rate of one click per minute or less. Figure 7.4 displays the number of speakers on the y-axis and number of total clicks on the x-axis.

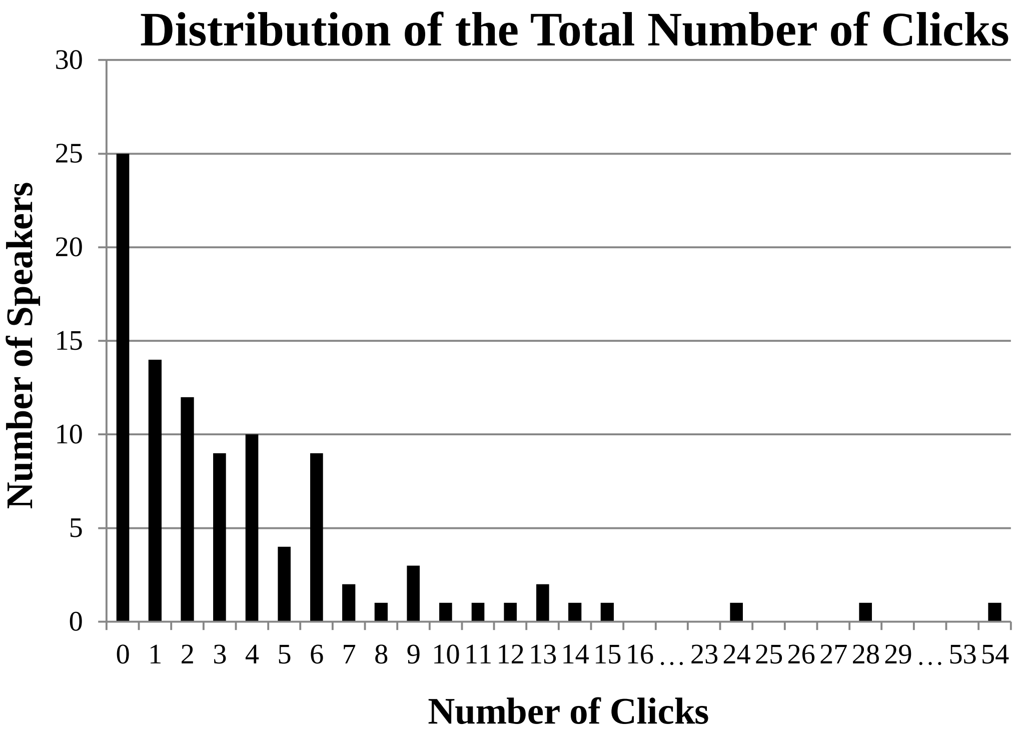


**Figure 7.4:** Distribution of click totals over five minutes of speech

Approximately 50% of speakers click only once, twice or not at all. And while the mean number of clicks for the group as a whole is 4.26 clicks over five minutes of net speech, this is highly skewed by three speakers who produce a very high number of clicks (24, 28, and 54). The mean number of clicks per speaker drops to 3.4 clicks when the three most extreme clickers are removed. Figure 7.5 presents the mean click rates in clicks per minute (clicks/min.), rather than as a cumulative number of clicks, as seen in Figure 7.4. The y-axis presents the number of speakers that fall within a given range and the x-axis depicts click rate in clicks per minute.

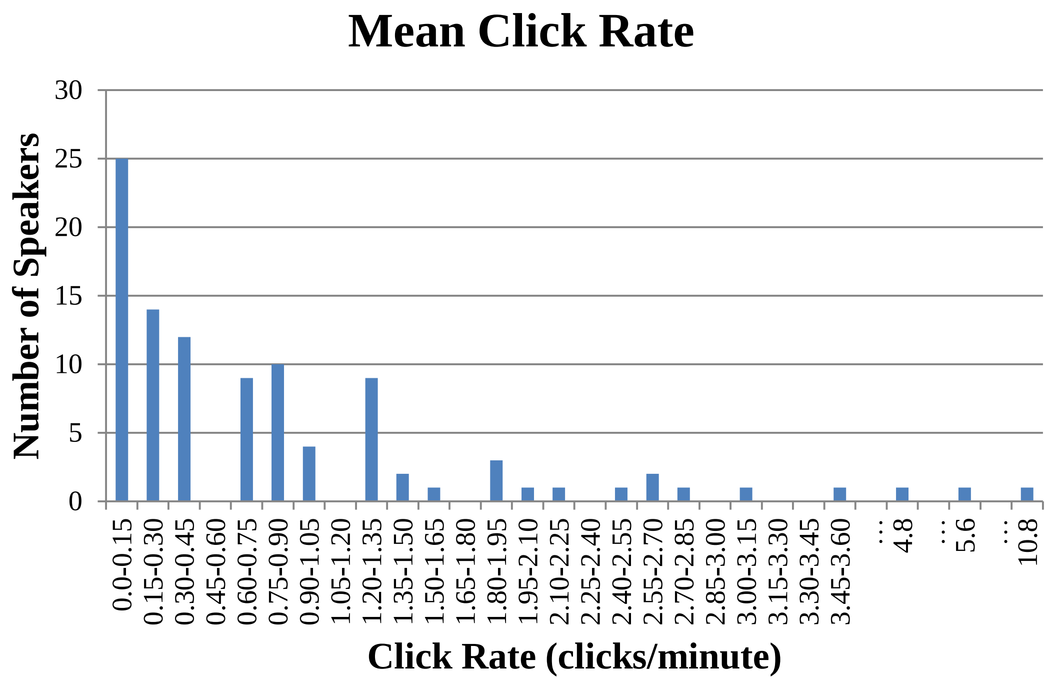


**Figure 7.5:** Distribution of click rate (clicks/minute) in DyViS population

There is an inevitable positive skew to the distribution of mean click rate in Figure 7.5. The mean click rate for the population is 0.88 clicks/min, with a range of 0.00 clicks/min to 10.8 clicks/min. The standard deviation of the means is 1.41 clicks/min. There are two suspected outliers at 3.00 clicks/min and 3.50 clicks/min. There are also three extreme outliers at 4.80 clicks/min, 5.60 clicks/min, and 10.80 clicks/min.

The cumulative distribution graph of mean click rates in Figure 7.6 shows the percentile at which a given click rate falls in relation to the population. The y-axis represents the cumulative proportion of the population, and the x-axis presents click rates in clicks per minute.

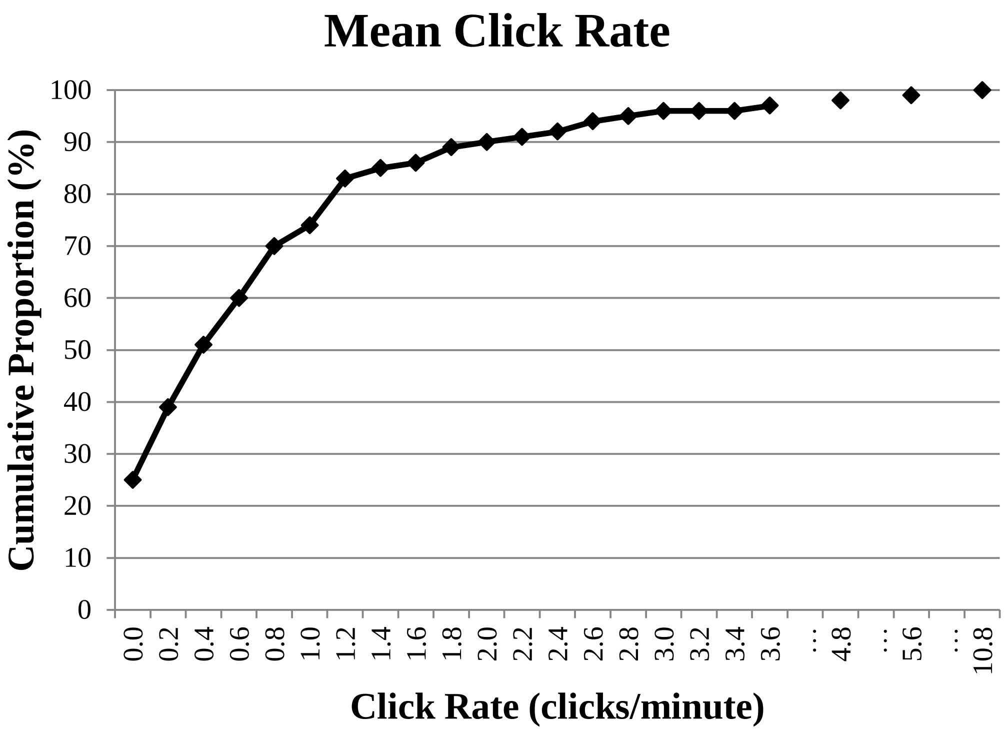


**Figure 7.6:** Cumulative percentages for click rate

The curve in Figure 7.6 starts at 25% for those with a click rate of 0, meaning that 25% of the DyViS population have no clicks present in their speech. From the first point at 25%, the curve is characterized as having an approximately logarithmic growth. Figure 7.6 shows that roughly 70% of the population have click rates at or below 0.8 clicks/min, and only 30% have larger click rates.

### 7.5.3.1 Discussion

Clicking, as a measure, has been shown to be highly sensitive to sample length (see Table 7.1), and it is not possible to specify a threshold sample duration for determining click rate, as the sample duration is dependent upon frequency of clicking. For example, to determine that someone has a click rate of, say, 0.2 clicks per minute, it would be necessary to have a sample five minutes in length, during which time the speaker clicks only once. However, to establish that someone had a click rate of, say 10 per minute, all one would need is one minute of speech or – indeed – less. This assumes, of course, that the clicks would be evenly distributed across time. And, as will be seen in the section below, such an assumption of intra-speaker stability is not supported by the data. For the present, however, it is noted that the low number of click totals for the majority of speakers makes the discrimination capacity of clicks difficult to establish. Nevertheless, there is *potential* for clicks to be a good discriminant for the handful of speakers who produce high click totals, if these speakers are relatively stable and consistent in their clicking behavior.

### 7.5.4 Results: Intra-Speaker Variation within an Interaction

The results for intra-speaker variation are presented in Figure 7.7. Speakers are represented on the x-axis and the click rates (clicks per minute) on the y-axis. A speaker's mean click rate is represented by a black dot, and the vertical bars indicate the range between the minimum and maximum click rate they attained in any individual minute of speech.

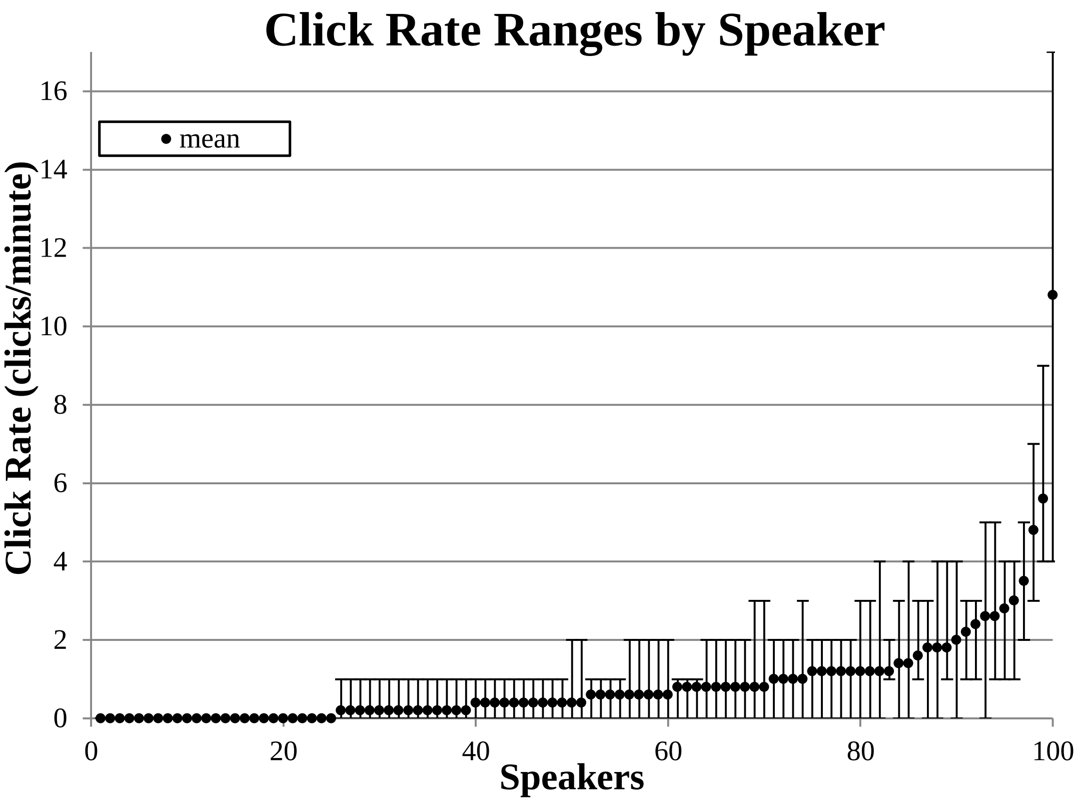


**Figure 7.7:** Mean and range of click rates across all speakers

It is clear from Figure 7.7 that intra-speaker stability generally increases as mean click rate increases, such that the higher-rate clickers have a greater range of variability across the individual minute blocks. Thus, even for those speakers for whom clicks might serve as a potentially discriminant feature, the clicks tend to occur in localized clusters rather than being evenly spread throughout the sample. This effectively means that in order to establish that someone has a high click rate, the analyst would need a relatively large amount of speech from him/her. In the forensic context, questioned recordings containing around one minute of net speech from the target speaker are not unusual. Obtaining five minutes of net speech is much less common**.** Thus, the possibility of using clicks as a discriminant feature in forensic casework, even for high-rate clickers, is quite limited.

There is a limited amount of data with which we can calculate variability. Nevertheless, inter-speaker variation is presented for the DyViS population. Caution must be exercised when interpreting the SD data. Figure 7.8 presents the distributions of standard deviations for the population.

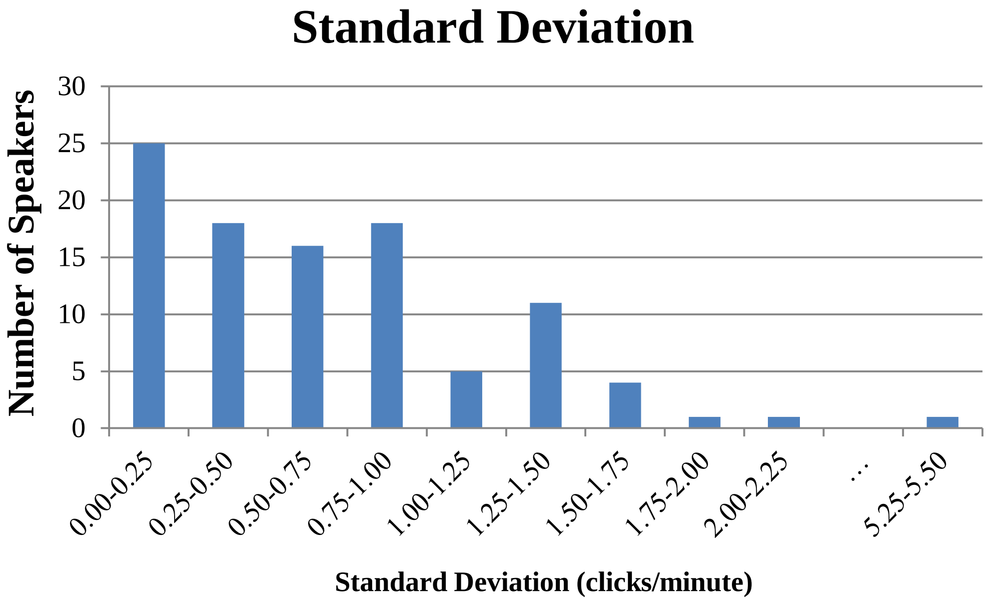


**Figure 7.8:** Distribution of standard deviation in click rate

Figure 7.8 has a positively skewed distribution, like that seen for mean click rate in Figure 7.5. There are two suspected outliers in the population (at 1.95 clicks/min and 2.07 clicks/min), and one extreme outlier at 5.40 clicks/min. The mean SD for click rate in the population is 0.69 clicks/min, with a range of 0 clicks/min to 5.40 clicks/min. The SD of the SDs for click rate is 0.70, which is actually higher than the mean, indicating a large spread in click rate values.

The cumulative distribution graph of SD for click rate is presented in Figure 7.9. The y-axis shows the cumulative proportion of the population with a SD at a given point, and the x-axis presents click rate in clicks per minute.

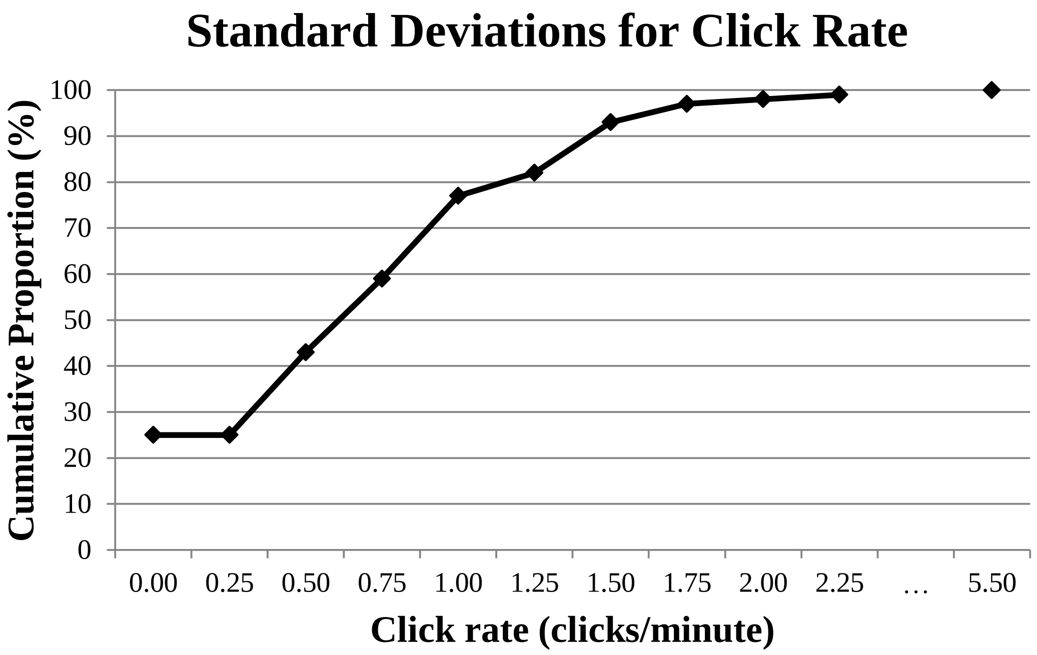


**Figure 7.9:** Cumulative percentages for standard deviation in click rate

The curve in Figure 7.9 is similar to the curve seen in Figure 7.6, but is slightly more gradient than logarithmic in its growth. The data in Figure 7.9 show that 25% of the speakers have SDs under 0.25 clicks/min., due to the 25% of speakers who do not click at all in their five minutes of net speech. The variance ratio for click rate is 4.06, which signifies that there is more variation between speakers than within speakers for click rate. A variance ratio of 4.06 is the highest that has been achieved for any parameter in the current thesis (articulation rate, long-term fundamental frequency, and long-term formant distributions). Despite a good variance ratio, caution has to be exercised, as it must be remembered that there were on average only five click tokens per speaker.

#### 7.5.4.1 Discussion

The sporadic distribution of clicks might be accounted for by the clustering of click opportunities (i.e. places where clicks can be used as discourse markers). There is no reason to assume that the need to express the affective meanings and perform the interaction management functions that clicks can fulfill should be evenly spread across time. A more detailed analysis might therefore address the question of the occurrence of clicks as a proportion of "click opportunities". Clustering of click opportunities can, of course, occur across interactions as well as within them, i.e. some types of conversation may well present more opportunities than others. For the present, however, another aspect of intra-speaker variation in clicking is examined, namely possible accommodation effects.

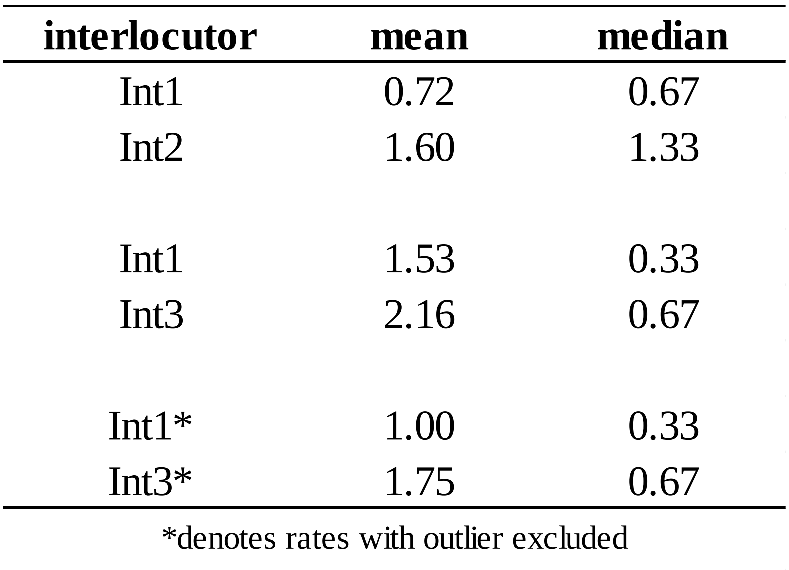
### 7.5.5 Results: Intra-Speaker Variation across Different Interactions

Accommodation, the tendency for speakers to adjust their speech towards that of their interlocutor, has been well documented in respect of a range of linguistic features (c.f. Giles, 1973; Giles and Ogay, 2007; Shepard, Giles and LePoire, 2001; Trudgill, 1981).

The click data considered so far were all drawn from the Task 2 recordings of the DyViS database, where each of the 100 subjects conversed with the same interlocutor, Int1. The recorded interviews that make up the Task 1 recordings involved two different interlocutors, Int2 and Int3, conversing with the 100 subjects. The further work reported in this section was triggered by the informal observation that the subjects appeared to be clicking more frequently in the Task 1 recordings when speaking with Int2 and Int3 than in the Task 2 recordings when speaking with Int1. This observation provided the motivation to undertake two further analyses: (a) establishing subjects' actual click rates in the Task 1 recordings relative to the Task 2 recordings, and (b) examining the click rates of Int2 and Int3 relative to Int1. The latter was undertaken with a view to determining whether any increase in subjects' clicking behavior might be accounted for by an interlocutor accommodation effect.

Fifty subjects were selected at random from the Task 1 recordings, 25 speaking with Int2 and 25 speaking with Int3.  As with the Task 2 sampling procedure, the first two minutes of the conversations were excluded from the analysis to allow for "settling-in time". Three minutes of net speech were extracted for each subject for comparison with an equivalent three-minute sample from the Task 2 recordings. Click rates were then compared across the two tasks. The comparisons showed that, although there was no statistically significant difference between the numbers of clickers *versus* non-clickers (using a chi-squared test, where $p$ = .7401 (Int1 to Int2) and $p$ = .0880 (Int1 to Int3)), clickers did show a marked increase in click rate when speaking to Int2 and Int3 over when speaking to Int1. The results are summarized in Table 7.2. The first column identifies the interlocutor, and the second and third columns present the mean and median click rates, respectively, for the given interlocutor.

**Table 7.2:** Summary of Speakers' Mean and Median Click Rates - Int1 versus Int2 and Int3



| interlocutor | mean | median |
|---|---|---|
| Int1 | 0.72 | 0.67 |
| Int2 | 1.60 | 1.33 |
| | | |
| Int1 | 1.53 | 0.33 |
| Int3 | 2.16 | 0.67 |
| | | |
| Int1* | 1.00 | 0.33 |
| Int3* | 1.75 | 0.67 |

*denotes rates with outlier excluded

The increase across Int1 to Int2 is significant at the 1% level (using a Wilcoxon signed rank test, *p* = .0034 and n = 25). That between Int1 to Int3 falls just short of significance at this level (1%), but achieves it if one speaker whose high click rate (speaker 07's click rate is 14.33 clicks/min for Task 1 and 12 clicks/min for Task 2) is excluded as an outlier (Wilcoxon signed rank test, *p* = .0076 and n = 24).

The actual changes – mean, minimum and maximum - for speakers are represented in Tables 7.3 and 7.4 (Int1 versus Int2 and Int1 versus Int3). The first column identifies the direction of change in click rate. The second column in Table 7.3 and 7.4 identifies the number of speakers with a given change in click rate, and columns three through five present the mean, minimum, and maximum changes in click rate for the group of speakers.

**Table 7.3:** Changes in click rate across speaker - Int1 versus Int2

| Δ (Int2-Int1) | Number of Speakers | Mean Δ | Minimum Δ | Maximum Δ |
|---|---|---|---|---|
| Increase | 15 | 1.601 | 0.003 | 3.003 |
| Same | 4 | — | — | — |
| Decrease | 6 | -0.334 | -0.003 | -0.670 |

**Table 7.4:** Changes in click rate across speaker - Int1 versus Int3

| Δ (Int3-Int1) | Number of Speakers | Mean Δ | Minimum Δ | Maximum Δ |
|---|---|---|---|---|
| Increase | 17 | 1.264 | -2.333 | 4.333 |
| Same | 0 | — | — | — |
| Decrease | 5 | 3.444 | -0.003 | -0.333 |

In attempting to account for the increases in click rates when subjects spoke to Int2 and Int3, click rates for Int2 and Int3 were calculated from three randomly-selected Task 1 recordings. The sampling procedure entailed extracting three minutes of net speech after the settling-in period, thus providing a total net sample of nine minutes for each interlocutor. For Int1 an equivalent portion of post-settling-in speech was extracted from the Task 2 recordings with the same three subjects selected for Int2 and Int3, thereby providing a total net sample of 18 minutes. The mean click rates for the three interlocutors are set out in Table 7.5. The first column identifies the interlocutor and the second column presents the mean click rate for the given interlocutor.

**Table 7.5:** Mean click rates of the three interlocutors

| interlocutor | mean click rate (clicks/minute) |
|---|---|
| Int1 | 1.44 |
| Int2 | 3.67 |
| Int3 | 4.56 |

Given the click rates established for subjects from the Task 2 recordings, Int1 might be seen as a relatively "average" clicker. Int2 and Int3, however, would be considered relatively high-rate clickers. In view of this, a plausible explanation for the increased click rates of the subjects when conversing with Int2 and Int3 would be that they are accommodating their clicking behavior towards that of their interlocutors. It is, of course, entirely possible that the accommodation effect is bilateral and that interviewers also adjust their click rates towards those of the subjects. The data to test this view are not available within the present study, however. Nor is it possible to assess whether interviewer gender is a factor[40]; it may or may not be significant that Int1 is a young male, while Int2 and Int3 are young women. An alternative, or indeed additional, explanation of the differences might be that the Task 1 interactions offer more clicking opportunities, these being mock police interviews in which the subjects are asked questions that might well have them searching for words in answering. However, this would not account for the relatively high click rates of Int2 and Int3, and although there are currently no formal findings to present on this, the clear impression is that there are no obvious differences amongst click opportunities.

## 7.6 Likelihood Ratios

The overriding limitation when analyzing the discriminatory power of clicks is the unfeasibility to calculate a numerical LR and evaluate the strength of evidence. The absence of an LR calculation for clicks is due entirely to the fact that a model does not currently exist with which it might be calculated.

[40] Accent may also be a factor, since only one of the interlocutors was also an SSBE speaker.

However, there are a number of mathematical procedures that can be used to arrive at a numerical LR. In forensics the different procedures are selected in relation to the characteristics of the data distributions. Aitken and Taroni (2004, p. 37) state that "for any particular type of evidence the distribution of the characteristic [parameter] is important. This is so that it may be possible to determine the rarity or otherwise of any particular observation." Therefore, it is important to use the model that best fits the distribution of data in order to represent the strength of evidence as accurately as possible.

Clicks are a particularly complicated form of speech evidence to work with when used to calculate numerical LRs, as they are discrete in nature. Aside from DNA profiling (which works with discrete data), there is a lack of methods when data are discrete rather than continuous. In forensic speech science, there has not been any LR research that has carried out a comprehensive analysis of discrete data. LR research in FSS has previously focused on continuous data (Gold and Hughes, 2013), for which it is possible to assume normality. Once an assumption of normality is made, "theory then allows for multivariate continuous data to be modeled using the means and covariances only" (Aitken and Gold, 2013, p. 148). However, for discrete data a description of the distribution as normal is not possible.

In this particular case, where there is a desire to calculate LRs for clicking rate in speakers, there are two main issues to consider when seeking how to model the data appropriately. The first is the possibility for each discrete data entry (e.g. the 5-minute recording) to have multiple levels of response (e.g. a click count for each minute in the recording). For example, in the present data, multiple levels of response are represented by the multiple click counts over a given amount of time. More specifically, Speaker A may have 5 minutes of net speech, where each minute of net speech yields an individual count (e.g. 0,0,1,2,0). The second issue is the correlation that exists between counts. Given that 25% of the population was found not to click at all over 5 minutes of net speech, it is apparent that correlations exist between counts, and these must be accounted for in a model.

The work in Aitken and Gold (2013) explores the issues and limitations involved in calculating LRs for discrete data. A Poisson distribution and bivariate Bernoulli model are proposed for evaluating clicks and any other discrete data that act in a similar way to clicks. The models proposed in Aitken and Gold are basic models; "however, they illustrate issues that need to be considered in the analysis of discrete data and provide a foundation on which other models may be built" (Aitken and Gold, 2013, p. 154). Likelihood ratios are provided in Aitken and Gold (2013, p. 153). However, they are based on a limited data set, whereby α and β (set distributions of the population) were not based on structural learning[41] but intuitive guesses about the population distribution. The LR results for clicks were between 0.30 and 3.35 (i.e. giving very limited evidence for support), which are small but "intuitively sensible" (Aitken and Gold, 2013, p. 154). More practical work is needed to further develop the models. However, it is hoped that further testing will also produce smaller LRs. Intuitively, this would align with there being a finite number of possible clicks produced over the course of a minute, high intra-speaker variation, and low inter-speaker variation.

[41] Structural learning makes decisions based on the data at hand, and uses those data to inform a given model/algorithm/framework (Porwal et al., 2013)

Given the limited strength of evidence reported for clicks in Aitken and Gold (2013), the lack of models for calculating click-rate LRs may not be all that devastating. This is due to the general lack of capacity of click rate to discriminate between speakers of English. As always, exceptions are to be made, however, for those individuals who lie towards the margins of the distribution curve and who can be classified as outliers with respect to click rate.

## 7.7 Conclusion

While it would be dangerous to generalize beyond the variety of English analyzed in this study, the view of those forensic practitioners surveyed in Chapter 3 who considered tongue clicking to be a highly discriminant feature of speaker behavior is largely unsupported by the present data for young male speakers of SSBE. Firstly, there is insufficient variation across the majority of speakers analyzed for the variable to provide a reliable index of speaker individuality. Secondly, even for the high-rate clickers who stand apart from the majority, there is within-conversation instability to the extent that one would need speech samples of a length seldom encountered in questioned forensic recordings in order to reliably establish an overall click rate. Thirdly, intra-speaker variation also occurs across interactions, apparently as a result of accommodation towards the clicking behavior of interlocutors. This suggests that rate of clicking, rather than being solely a property of an individual's speech production practices, might usefully be viewed as resulting from an interaction between speaker and interlocutor. The question remains, then, of whether it is worth considering clicking at all when conducting speaker comparison casework. In spite of these findings, it is suggested that, in certain cases, it may well be. Studies such as those of Wright (2007; 2011a; 2011b) and Ogden (2013) on the interactional functions of clicks, as well as the general observations of phoneticians on the functions of clicks in conveying attitudinal and affective meanings, provide normative data and descriptions. For this reason, these studies allow forensic practitioners to assess the speech samples they examine for the occurrence of non-normative, i.e. idiosyncratic, usage. Such occurrences may be of assistance in the comparison task, and in this respect forensic phoneticians are indebted to their non-forensic counterparts for providing valuable resources. This is, in fact, just a further instance of a more general indebtedness of the forensic speech community to work in mainstream academic research in linguistics and phonetics. As noted in French and Stevens (2013), sociophoneticians and dialectologists have provided normative descriptions of language varieties that serve as backcloths for the evaluation of findings in speaker comparison cases.

Unless it were to transpire that patterns of clicking behavior are different for other varieties of English or (for example) differ in accordance with speaker age or gender - and nothing has been found in the sociolinguistic literature on English to support that view - the mere comparison of click rates across samples is in the overwhelming majority of cases unlikely to advance the speaker comparison task, for the reasons outlined above.

# Chapter 8: Overall Likelihood Ratios

## 8.1 Introduction

"The whole is greater than the sum of the parts." - Aristotle

The survey of FSC practices (Chapter 3) revealed that for the vast majority of expert forensic phoneticians, it is the overall combination of parameters that they consider crucial in discriminating between speakers (despite some parameters having greater weight than others). For this reason, the current chapter addresses the issue of combining parameters for speaker discrimination through empirical testing.

The combination of phonetic, linguistic, and non-linguistic parameters in an FSC has traditionally been carried out by experts through implicit 'mental' calculations. That is to say, an expert creates a mental representation of the properties of an individual's speech and makes a judgment about the likelihood that the speakers in the suspect and criminal samples are the same person (based on the combined weight of the evidence). The process by which an expert 'mentally' combines parameters to arrive at a conclusion is not transparent. As such, it has been argued that different experts will weigh certain parameters more highly than others, based purely on personal opinions (Rose and Morrison, 2009). For this reason, the traditional method of parameter combination in FSCs is highly subjective and is difficult to replicate.

Bayes' theorem, on the other hand, offers a more explicit and transparent alternative for the combination of parameters. A simple combination procedure, known as 'naïve Bayes' (Kononenko, 1990), involves multiplying the individual

LRs (or equivalently, the addition of individual LLRs) assuming that there is mutual independence between parameters (i.e. the parameters are not correlated). The combination of correlated parameters is a problem for the LR framework, because unless parameters are mutually independent there is a risk of over-estimating the strength of evidence by considering the same parameter more than once. Alternative methods such as logistic-regression fusion, MVKD, and Bayesian Networks  have been put forward to circumvent the problem whilst maintaining a Bayesian approach (Aitken and Lucy, 2004; Brümmer et al., 2007; Gonzalez-Rodriguez et al., 2007; Morrison et al., 2010). However, given the lack of appropriate testing, it is unclear whether logistic-regression fusion or MVKD adequately take account of correlations in the data.

The aim of the present chapter is to amalgamate the individual speech parameters from Chapters 4-7 into a complete system[42], whereby discriminant power, strength of evidence, and validity can be tested for all analyzed speech parameters in combination. Previous research has developed methods to facilitate the combination of individual speech parameters into some form of a combined system. However, the blend of approaches taken in this chapter has never been used before. The chapter begins by exploring the existing relationships between LTFD, AR, F0, and click rate to check for potential correlations. The correlation coefficients are then used to inform appropriate combination methods given the (in)dependencies that exist amongst the given

[42] A system is defined by the Oxford English Dictionary (http://www.oxforddictionaries.com/definition/english/system) as "a set of things working together as parts of a mechanism or an interconnecting network; a complex whole". The term 'system' is often used in the ASR literature to refer to a 'complex whole' that is well-suited to providing a response to two competing hypotheses being tested (the evidence given the prosecutor's hypothesis divided by the evidence given the defense's hypothesis). The term *system* is extended in this thesis to both individual speech parameters and speech parameters in combination that can also provide the basis for an evaluation of the two competing hypotheses.

parameters. After the combination of individual parameters into a complete working system, the discriminant ability, strength of evidence, and validity are tested for the combined parameters. The integration of methodological approaches employed in this chapter is a first for calculating overall likelihood ratios (OLRs). This approach is intended to demonstrate how an analyst would go about using these methods in order to avoid an over- or under-estimation of the strength of evidence for calculating OLRs for the data under scrutiny.

## 8.2 Literature Review

The combination of forensic speech evidence under a numerical LR framework has received a reasonable amount of attention in the literature (Alderman, 2004; Kinoshita, 2002; Rose et al., 2003; Rose et al., 2004; Rose and Winter, 2010). The main focus of this earlier research is the issue of combining parameters that are potentially correlated. Early studies evaluating traditional linguistic phonetic parameters often recognized this problem but did nothing to try to ameliorate it. For example, Kinoshita (2002) used naïve Bayes to combine LRs based on the best-performing set of formant predictors from /m/, /ʃ/, and a set of short vowels into a single expression of posterior probability. Similarly, Alderman (2004) generated an OLR from different vowel formant predictors using naïve Bayes in order to compare the speaker-discriminatory performance of different combinations of parameters (and individual features of parameters).

Rose et al. (2003) displayed a more overt awareness of the issues surrounding correlation within and between parameters. In their study, they compared the discriminatory performance of formants using segmental cepstra from /ɔː ɕ ɴ/ in Japanese. Linear regression was applied to the parameters to assess the degree of correlation only within individual parameters (i.e. the formants). The individual LRs were combined into an OLR using an assumption of independence, although between-parameter correlation was never explicitly tested because “it was assumed that, given the very different phonetic nature of the three segments used, there was unlikely to be much correlation between all but their highest formants” (Rose et al., 2003, p. 195). Rose et al. (2003) make a good attempt at accounting for within-parameter correlation, but fail to go one step further to test the between-parameter correlations. Phonetic theory would predict that the parameters are not correlated. However, without further testing, correlations may go unexposed (and unrealized). Rose et al. (2003) also note that linguistic theory leads them to believe that the higher formants may be correlated, yet nothing is done to account for it. Therefore, it is probable that the results produced for the study were over- or under-estimations of the strength of evidence.

The development of LR modeling techniques has brought with it the capability of dealing more appropriately with the complexities of correlation. Aitken and Lucy’s (2004) MVKD formula treats the set of data from which LRs are computed as multivariate data, and as such is able to account for within-segment correlation. Rose et al. (2004) investigated the comparative performance of the multivariate LR approach and the naïve Bayes assumption of independence. The naïve Bayes approach was shown to overestimate the strength of SS and DS LRs compared with the more conservative MVKD model. The proportion of errors was also better when independence was assumed, which led Rose et al. (2004) to conclude that “the ‘correct’ formula is still not exploiting all the discriminability in the speech data and (as such) the Idiot's approach [naïve Bayes] is still preferable" (Rose et al., 2004, p. 496). However, the study fails to discuss the fact that naïve Bayes produces misrepresentative estimates of the strength of evidence when parameters are correlated, which could lead to a miscarriage of justice in a real case.

Another recently-adopted technique to account for potential correlation between phonetic-linguistic parameters in LR-based FSC is the logistic regression fusion approach. Fusion is a form of "back-end processing" (Rose and Winter, 2010, p. 42) which attaches weights to parameters based on correlations between LRs from individual parameters. This contrasts with "front-end processing", which considers correlations in the raw data. Fusion was developed within the field of ASR (Brümmer et al., 2007; Gonzalez-Rodriguez et al., 2007; Ramos Castro, 2007) and has since been applied in a number of studies using traditional phonetic parameters (Morrison, 2009; Morrison et al., 2010; Rose, 2010b; Rose, 2011) leading Rose and Winter (2010, p. 42) to claim that fusion is one of the "main advances" to have emerged from automatic methods.

Fusion is currently the only alternative to a naïve Bayes approach for LR-based forensic phonetic analysis. However, there are a number of potential problems with fusion. Firstly, back-end processing, as the name suggests, deals with correlations after the generation of numerical LRs has been performed. Therefore, as suggested by Rose, "it is … possible … that two segments which are not correlated by virtue of their internal structure and which therefore should be naively combined, nevertheless have LRs which do correlate" (Rose, 2010, p. 32). Equally, the reverse is possible, whereby correlated parameters generate non-correlated LRs. More broadly, there is also an issue of efficiency. Since fusion is implemented after the generation of LRs, the original analysis may unnecessarily include a number of highly-correlated parameters, which when combined provide a limited strength of evidence.

## 8.3 Data

The present study does not introduce any new data, as it works with the data presented in Chapters 4-7. The parameters under consideration are mean long-term formant frequency distributions (for F1-4), mean articulation rate, long-term mean fundamental frequency, and click rate.

## 8.4 Correlations

This section considers potential correlations that exist within and between parameters. Correlations are calculated for the speakers as a group, as opposed to individual speakers. Therefore, it could be the case that the correlations found for the group of 100 speakers do not exhibit the same patterns as those calculated for an individual speaker.

### 8.4.1 Methodology

Correlations were calculated to identify potential relationships or mutual independencies within and between parameters. Two groups of correlations were calculated for the data: those within LTFD (i.e. LTFD1, LTFD2, LTFD3, LTFD4) and those between parameters (LTFD1-4, AR, F0, click rate). The formants within LTFD are treated as individual parameters for correlation testing, given that phonetic theory has established that individual formant measurements represent different physiological aspects of a vowel (e.g. frontness/backness of the tongue, height of the tongue, voice quality; Ladefoged, 2006; Laver, 1994). Correlation coefficients were calculated for individual LTFD comparisons by selecting two LTFD measurements at a given data point, resulting in hundreds of data points per formant comparison. Calculating correlation coefficients between parameters required a single data point per person, so a mean value was calculated for each speaker's LTFD1-4, AR, and F0 (i.e. three separate means). The data for click rate already existed as a single data point for each speaker, so no additional mean calculations were required.

All correlations in this section were calculated using Spearman's rank correlation coefficient. This method was preferred over Pearson correlation coefficients, as the latter assesses how well the relationship between two variables can be described using a monotonic function. The Pearson correlation coefficient is calculated on the assumption that the relationship between two variables is linear. Because the pair-wise relationships between variables under consideration are not known (nor can they be assumed to be linear), the Spearman's rank correlation coefficient was the logical choice.

MatLab (version R2012a) was used to create scatterplots and to calculate the correlation coefficients for all pairs of parameters. Table 8.1 presents all six possible pairing combinations for the LTFD parameter, and Table 8.2 presents all 15 possible pairing combinations between parameters.

**Table 8.1:** Formant pairings within LTFD

| Parameter 1 | Parameter 2 |
|---|---|
| *Long-term Formant Distributions* | |
| LTFD1 | LTFD2 |
| LTFD1 | LTFD3 |
| LTFD1 | LTFD4 |
| LTFD2 | LTFD3 |
| LTFD2 | LTFD4 |
| LTFD3 | LTFD4 |

**Table 8.2:** Between-parameter pairings

| Parameter 1 | Parameter 2 |
|---|---|
| LTFD1 | Mean F0 |
| LTFD1 | AR |
| LTFD1 | Click Rate |
| LTFD2 | Mean F0 |
| LTFD2 | AR |
| LTFD2 | Click Rate |
| LTFD3 | Mean F0 |
| LTFD3 | AR |
| LTFD3 | Click Rate |
| LTFD4 | Mean F0 |
| LTFD4 | AR |
| LTFD4 | Click Rate |
| Mean F0 | AR |
| Mean F0 | Click Rate |
| AR | Click Rate |

Tables 8.1 and 8.2 are complete lists of all 21 parameter pairings. The first and second columns simply identify the parameters that are being compared against each other.

The point at which two parameters can be deemed to be correlated is a matter of subjective judgment, in that there is no specific correlation coefficient that explicitly signifies dependence between two parameters. The decision of independence is made by the expert, which in turn can result in different opinions regarding the threshold at which correlations are implied (ultimately, this can cause variation in the LR results). For the purpose of this study, correlations were considered through structural learning (see § 7.6), which is informed by the data rather than theoretical considerations. Final correlation judgments were made by me after examining scatterplots in conjunction with correlation coefficients for each pair-wise comparison. My judgments relating to correlations were also confirmed by a forensic statistician (Marjan Sjerps, p.c.).

## 8.4.2 Within-Parameter Correlation Results

The scatterplots for all pair-wise comparisons within LTFD are presented in Figures 8.1 - 8.6. The y-axis presents one LTFD parameter, while the x-axis represents another.

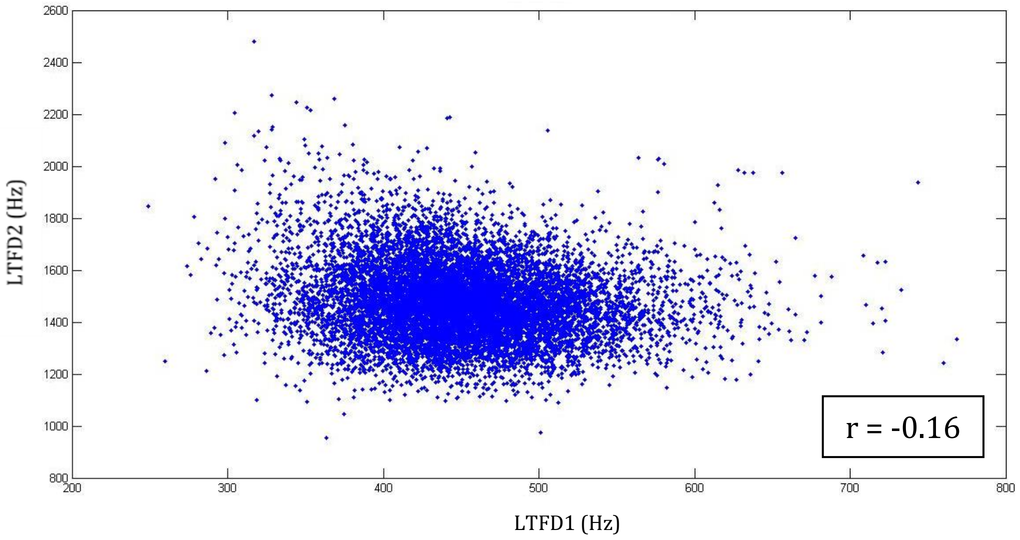


**Figure 8.1:** LTFD1 versus LTFD2

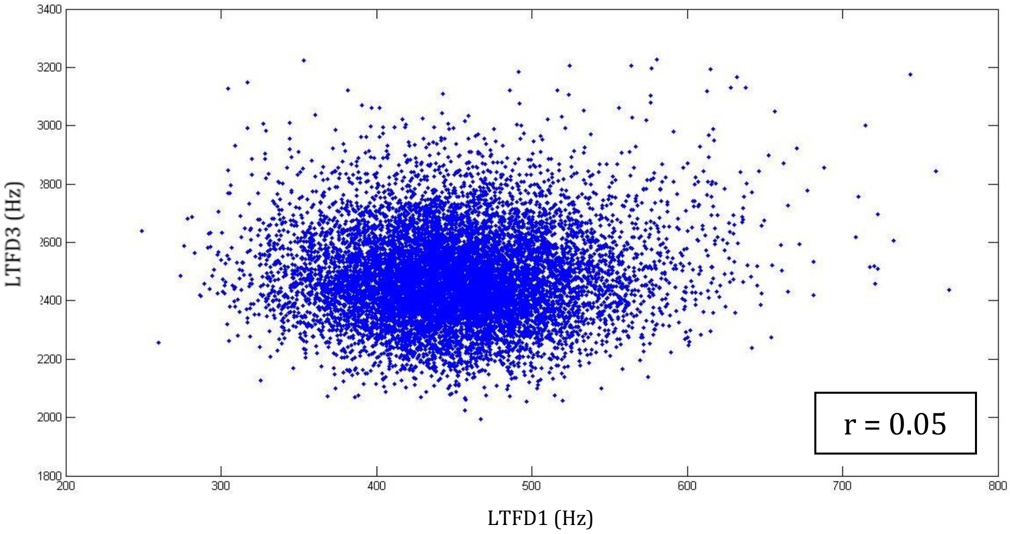


**Figure 8.2:** LTFD1 versus LTFD3

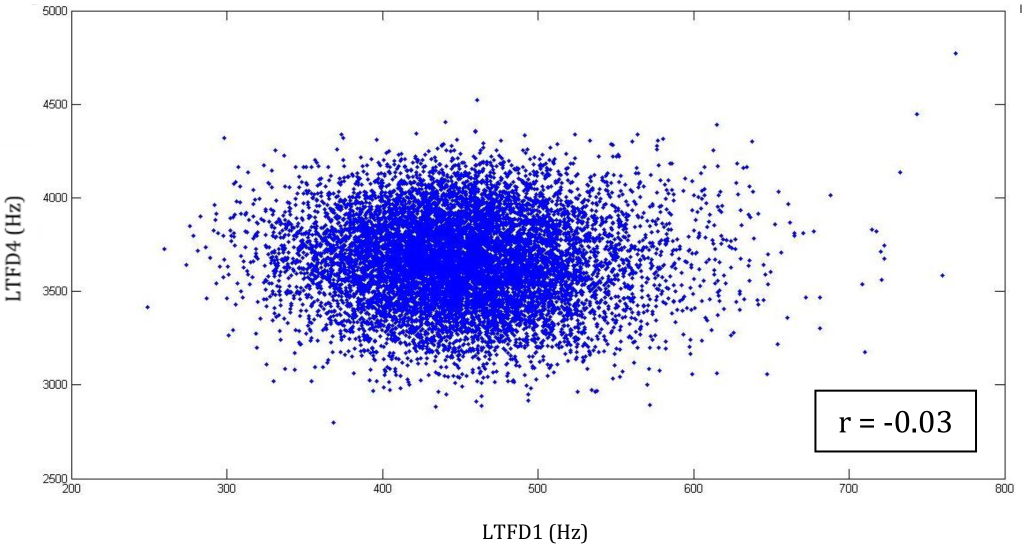


**Figure 8.3:** LTFD1 versus LTFD4

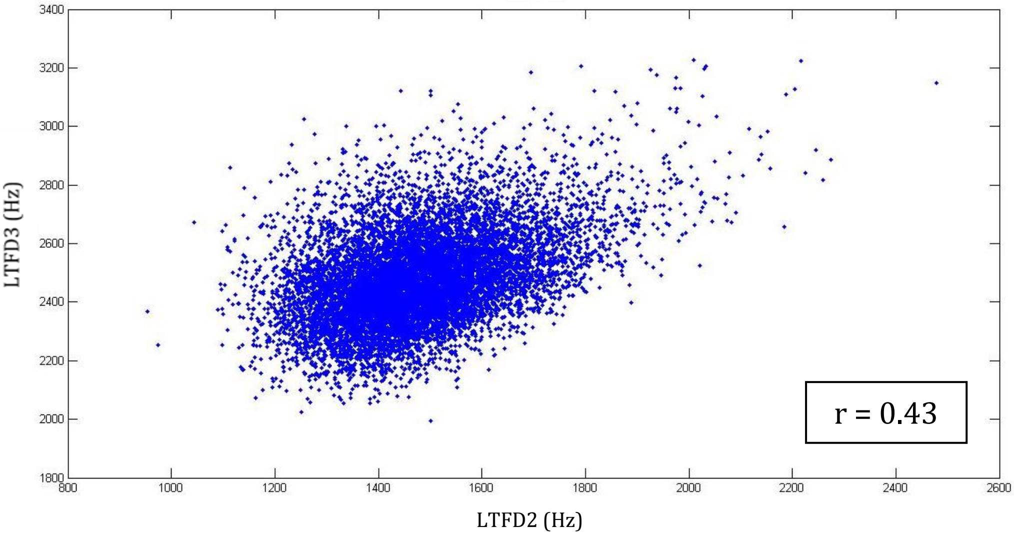


**Figure 8.4:** LTFD2 versus LTFD3

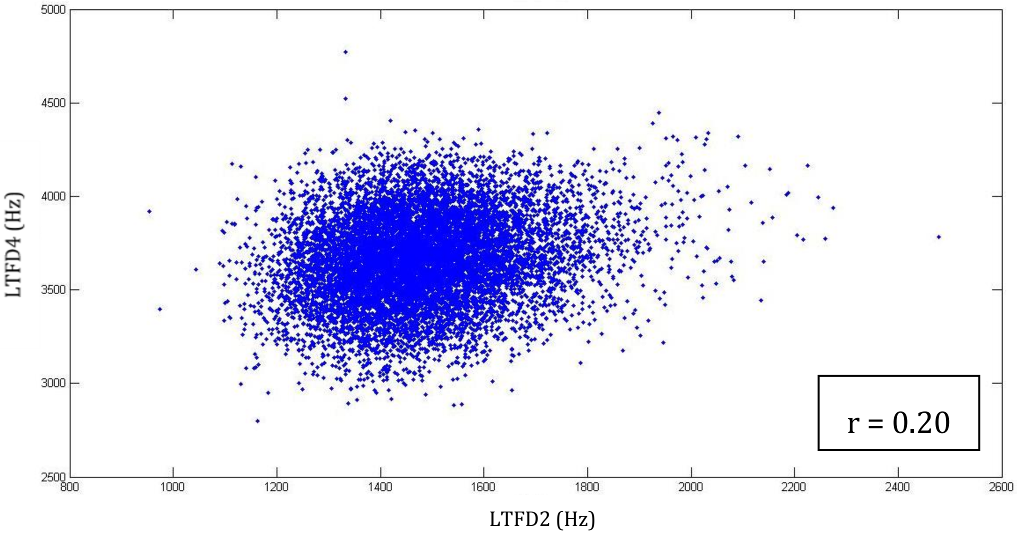


**Figure 8.5:** LTFD2 versus LTFD4

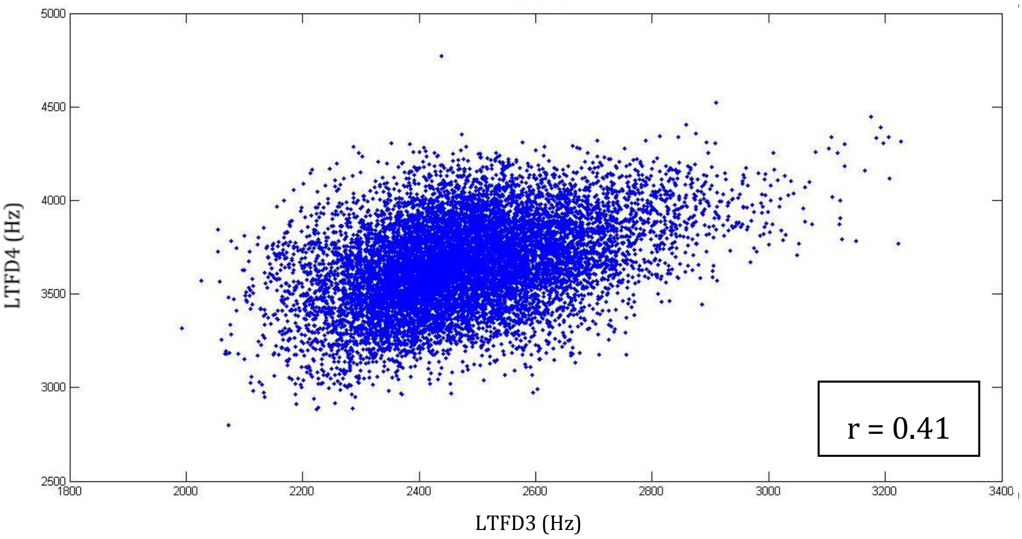


**Figure 8.6:** LTFD3 versus LTFD4

The scatterplots in Figures 8.1 - 8.3 do not exhibit any strong relationships between the variables, and graphically suggest that there are no correlations. Figures 8.4 and 8.6 are characterized as having moderate positive correlations, while Figure 8.5 has a slightly weaker positive correlation. The correlation present in Figure 8.5 (LTFD2 vs. LTFD4) is most likely representative of indirect correlation, given that LTFD2 correlates with LTFD 3, and LTFD3 correlates with LTFD4.

The correlation coefficients for within-LTFD comparisons are presented in Table 8.3. The intersection of a column and row indicates a given comparison, and the value within the box is Spearman's rank correlation coefficient (*r*). A value closer to 1 or -1 suggests that two parameters are correlated, while a value close to 0 suggests the two parameters are not correlated.

**Table 8.3:** Correlation coefficients within LTFD

| | LTFD1 | LTFD2 | LTFD3 | LTFD4 |
|---|---|---|---|---|
| **LTFD1** | 1.00 | -0.16 | 0.05 | -0.03 |
| **LTFD2** | | 1.00 | 0.43 | 0.20 |
| **LTFD3** | | | 1.00 | 0.41 |
| **LTFD4** | | | | 1.00 |

Based on the results seen in Figure 8.1 - 8.6 and Table 8.3, an informed judgment can be made with regard to which parameters appear to be correlated within LTFD1-4. The results suggest that LTFD2 is correlated with LTFD3, LTFD3 is correlated with LTFD4, and LTFD2 is indirectly correlated with LTFD4 (they have a transitive relationship by way of LTFD3; this also referred to as a partial correlation). LTFD1 and LTFD2 have a correlation coefficient of -0.16. However, this correlation was not deemed to be significant (*r* is less than 0.25; confirmation also given by Marjan Sjerps, p.c.). Therefore, LTFD1 is reasoned to be independent from LTFD2-4.

## 8.4.3 Between-Parameter Correlation Results

The scatterplots for all pair-wise comparisons between parameters are presented in Figures 8.7 - 8.12. The y-axis represents the first parameter, and the x-axis represents the second parameter.

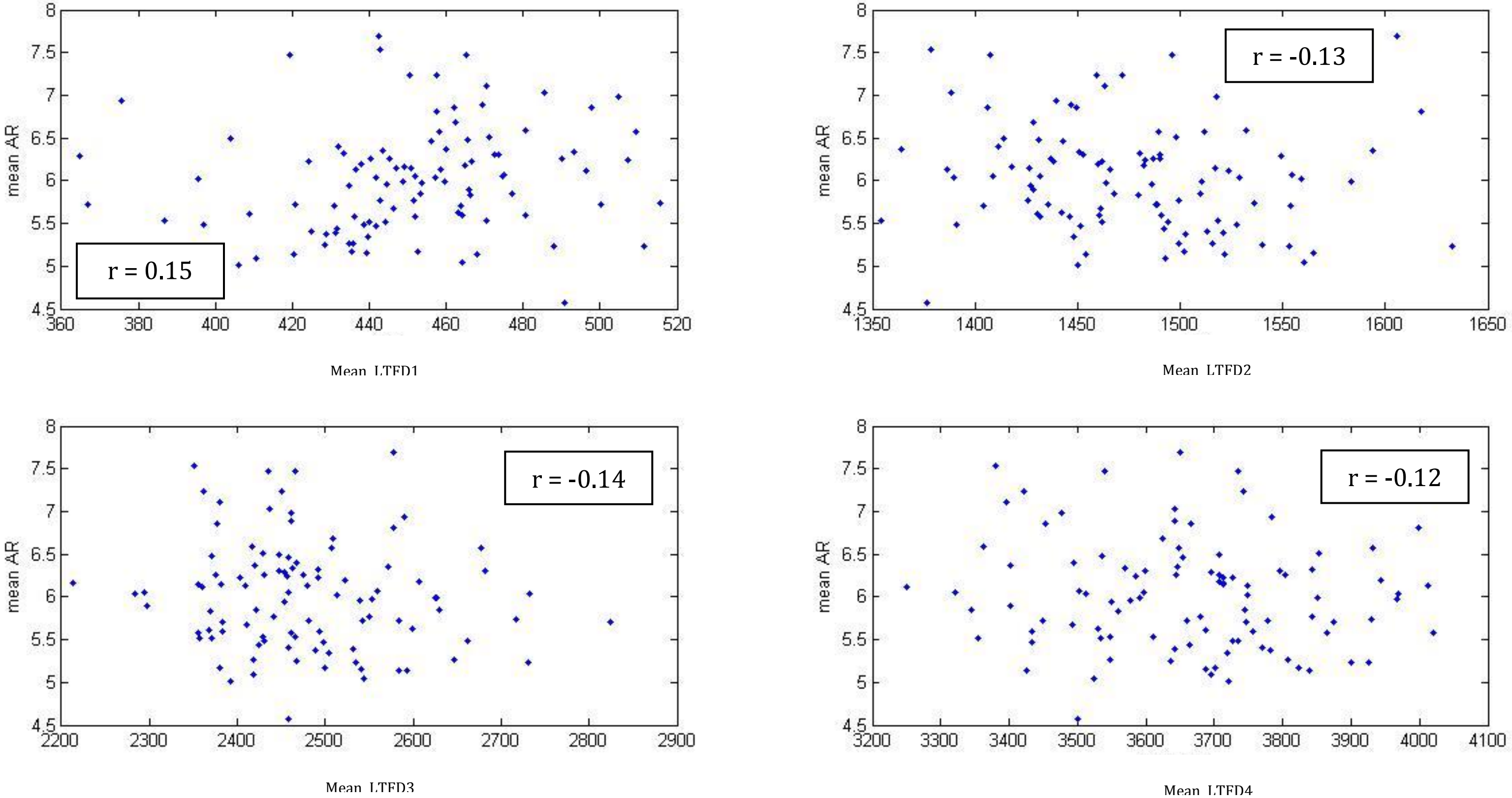


**Figure 8.7:** Mean AR versus LTFD1-4

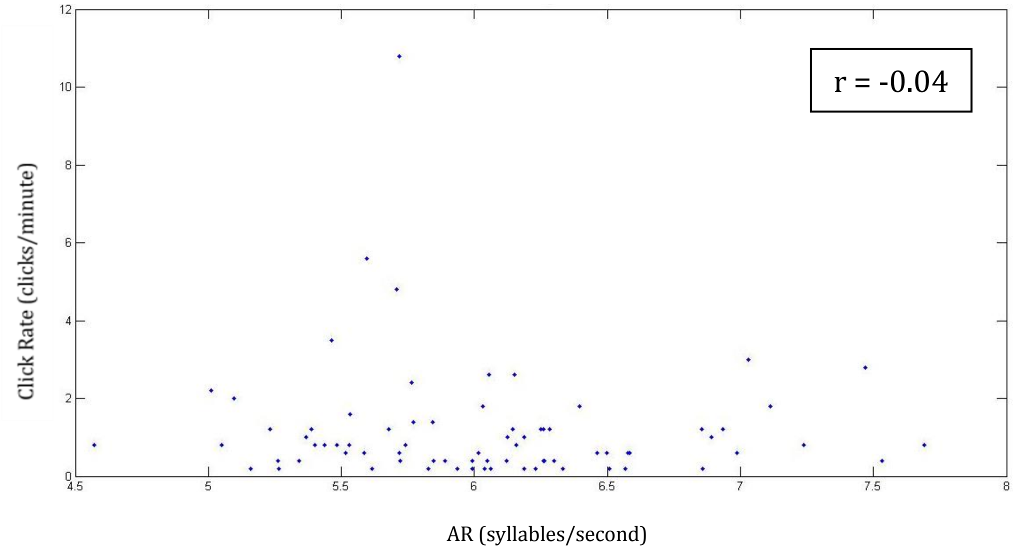


**Figure 8.8:** Click rate versus mean AR

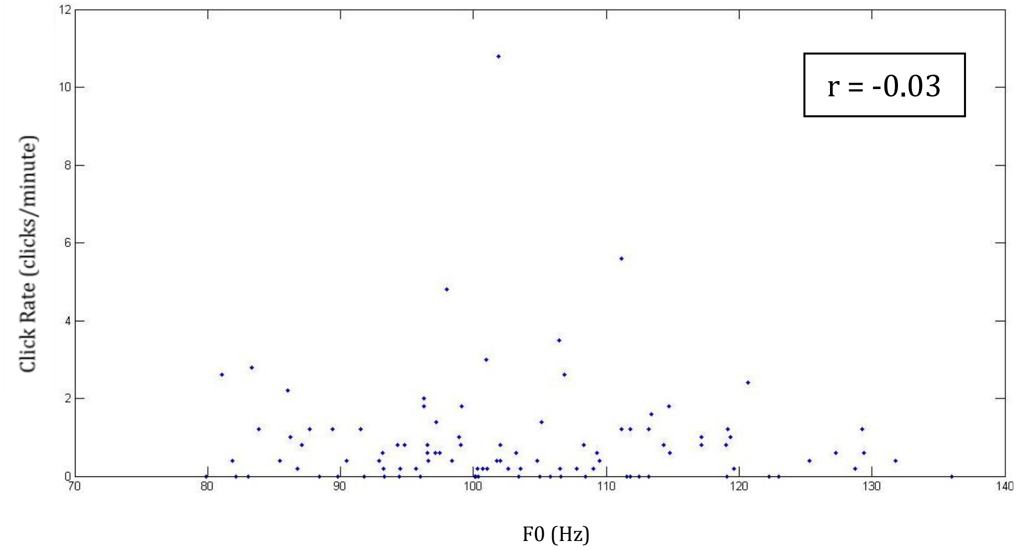


**Figure 8.9:** Click rate versus mean F0

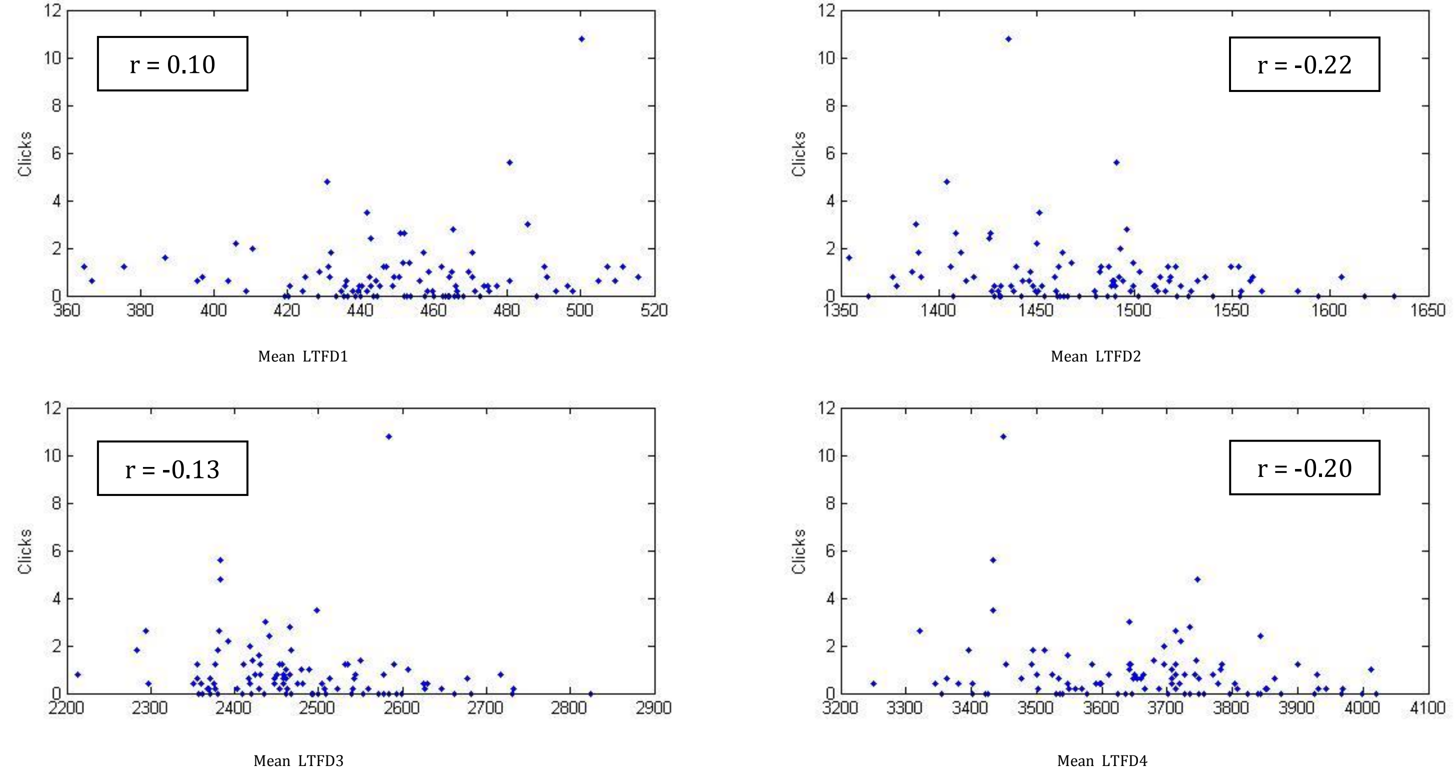


**Figure 8.10:** Click rate versus LTFD1-4

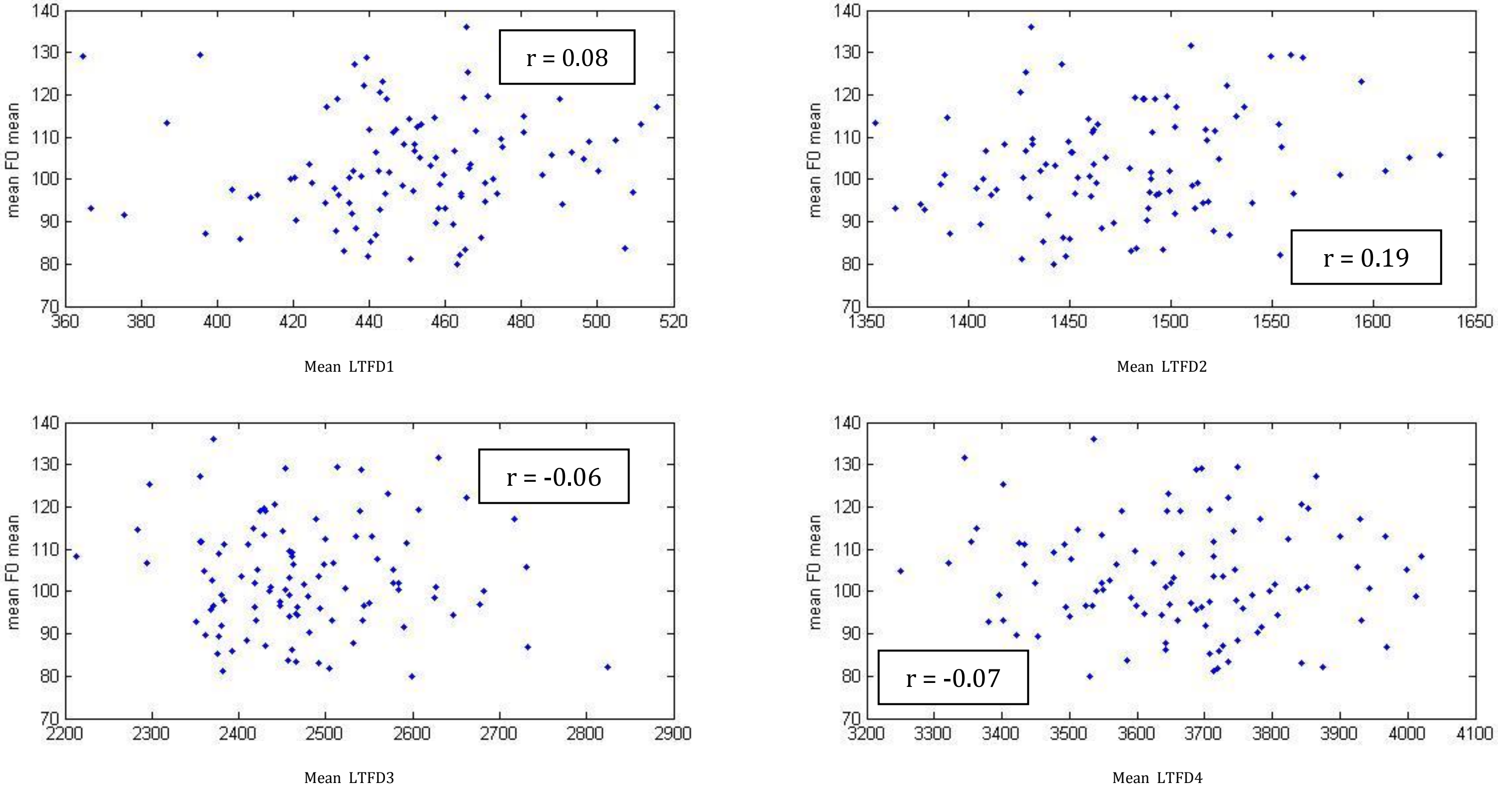


**Figure 8.11:** Mean F0 versus LTFD1-4

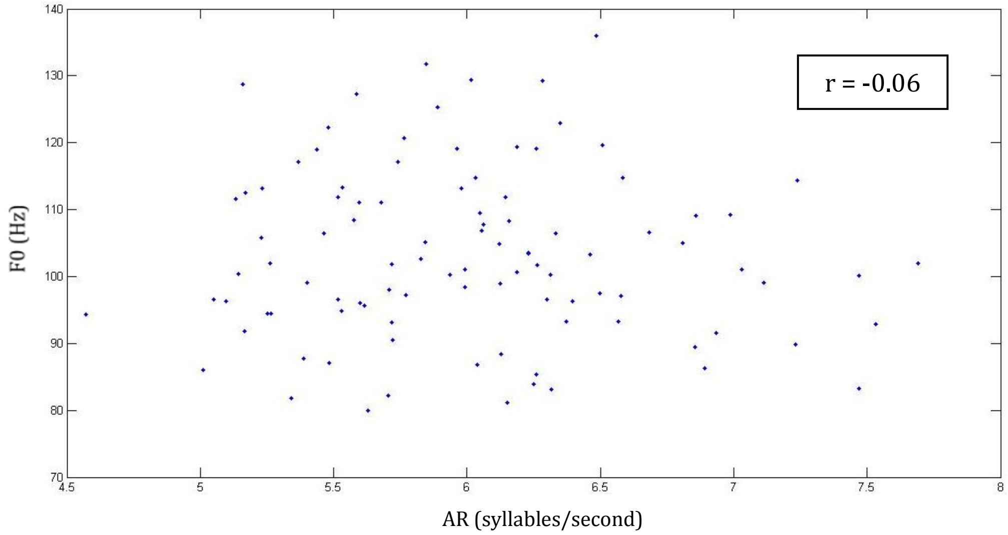


**Figure 8.12:** Mean F0 versus mean AR

The scatterplots in Figures 8.7 - 8.12 do not exhibit any signs of strong (or even moderate) correlations between any of the parameter pairings. Because the scatterplots have a limited number of data points (only 100 in this case) compared with the number of data points for Figures 8.1 - 8.6, the calculation of correlation coefficients is necessary (as was also seen in Table 8.3) to quantify the levels of correlation between parameters.

The results of the correlation coefficients are presented in Table 8.4.

**Table 8.4:** Correlation coefficients within- and between-parameters

| | LTFD1 | LTFD2 | LTFD3 | LTFD4 | AR | F0 | Click Rate |
|---|---|---|---|---|---|---|---|
| **LTFD1** | 1.00 | -0.16 | 0.05 | -0.03 | 0.15 | 0.08 | 0.10 |
| **LTFD2** | | 1.00 | 0.43 | 0.20 | -0.13 | 0.19 | -0.22 |
| **LTFD3** | | | 1.00 | 0.41 | -0.14 | -0.06 | -0.13 |
| **LTFD4** | | | | 1.00 | -0.12 | -0.07 | -0.20 |
| **AR** | | | | | 1.00 | -0.06 | -0.04 |
| **F0** | | | | | | 1.00 | -0.03 |
| **Click Rate** | | | | | | | 1.00 |

In terms of between-parameter correlations, Table 8.4 does not present any new strong correlations (those within LTFD have already been discussed in § 8.4.2). The strongest relationships found between parameters are those for LTFD2 vs. click rate (-0.22), LTFD4 vs. click rate (-0.20), and LTFD2 vs. mean F0 (0.19). Linguistics literature and phonetic theory do not give any reason to lead one to believe that these parameters should be related to one another[43]. Therefore, the very weak correlations seen in Table 8.4 have most likely happened by chance. Correlation does not imply causality, and these three cases appear to be good examples of this.

### 8.4.3.1 Discussion

Based on the results seen in Figures 8.7-8.12 and Table 8.4, an informed judgment can be made with respect to the parameter correlations for the data set. The results suggest that there is no parameter correlation between LTFD, mean AR, mean F0, and click rate (confirmation given by Marjan Sjerps, p.c.). As such, these parameters are deemed to be mutually independent from one another for this particular data set.

## 8.5 Overall Likelihood Ratios

Based on the interdependencies and conditional dependencies found in § 8.4, OLRs can be calculated for the system. A model does not currently exist with

[43] However, it is possible that F0 and F2 could be related. High F2 values are associated with tongue fronting, and as the tongue body fronts, it pulls on the hyoid bone, from which the larynx is suspended. Laryngeal tension of this sort would promote higher F0, because of tension on the vocal folds. Therefore, it is possible that one might anticipate a correlation between high F2 values and high F0 values.

which calculate numerical LRs for click rate, and therefore the OLRs calculated in this section exclude this parameter from analysis.

### 8.5.1 Methodology

The OLRs presented in this section were calculated in multiple stages. Individual LRs were calculated for LTFD1, AR, and F0 using a MatLab implementation of Aitken and Lucy's (2004) MVKD formula (Morrison, 2007), and a separate LR was calculated for LTFD2-4 together using the same MVKD formula. This was done in order for the algorithm to take into account the correlations that exist between these three parameters (LTFD2-4). An intrinsic methodology, whereby the test and the reference speakers came from the same population of 100 speakers, was used for all LR calculations. Speakers 1-50 were used as the test speakers, while speakers 51-100 served as the reference speakers.

The results from the individual LRs and the LR from LTFD2-4 were then multiplied together following Naïve Bayes (given that § 8.4 demonstrated that AR, F0, LTFD1, and LTFD2-4 were independent of one another) to form a complete system. Additional variations of the system were also computed in the same manner as for the complete system, whereby the LR for LTFD2-4 is always calculated together (in the MVKD formula) and multiplied by the other individual LRs in different combinations. A MatLab script[44] was then used to calculate basic statistics, EER, and Cllr for the OLR system and variations on this system.

[44] This script was developed by Phil Harrison of J P French Associates.

Logistic-regression calibration was also applied to the complete system in two different orders using a MatLab script[45]. Logistic-regression calibration (see § 8.5.3 for further discussion) was applied in the first instance to individual parameters before combination, and applied in the second instance to the complete system after the parameters had been combined in order to compare the effectiveness of the calibration (in terms of EER and Cllr).

### 8.5.2 Overall Likelihood Ratio Results: Uncalibrated

The results of the OLR for the complete system are provided in Table 8.5. The complete system is composed of LTFD1, LTFD2-4, F0 (mean and standard deviation), and AR (mean). The first column in Table 8.5 presents the comparison type (SS or DS pairs), followed by the percentage of correct pairs, mean LLR, minimum LLR, and max LLR. The final two columns report on the complete system's validity, where the sixth column provides the EER and the final column presents Cllr.

**Table 8.5:** Summary of LR-based discrimination for the complete system (100 speakers)

| Comparison | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **Complete System SS** | 92.00 | 5.673 | -3.082 | 7.316 | .0607 | .3793 |
| **Complete System DS** | 93.27 | 1.560 | -infinity | 3.963 | | |

Table 8.5 shows that the combination of all parameters into the complete system provides an EER of 0.0607, and a Cllr of 0.3793. It appears that the complete system is good at identifying SS pairs, and slightly better at identifying DS pairs. The strength of evidence that the system offers is considerably stronger than that seen in the tests reported in Chapters 4-6. Figure 8.13

[45] This script was created by Niko Brümmer, modified by Geoffrey Morrison, and edited by Vincent Hughes.

presents the Tippett plot of the complete system, and Figure 8.14 is a zoomed-in version of Figure 8.13.

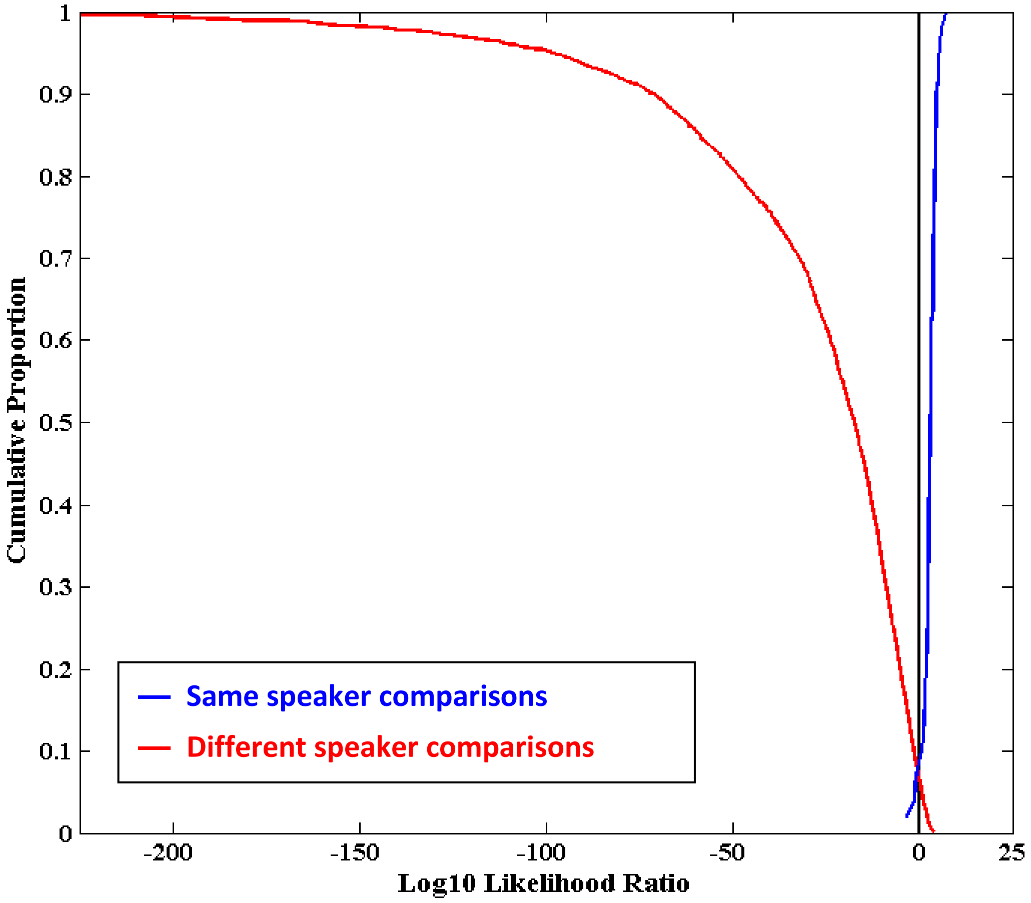


**Figure 8.13:** Tippett plot of the complete system

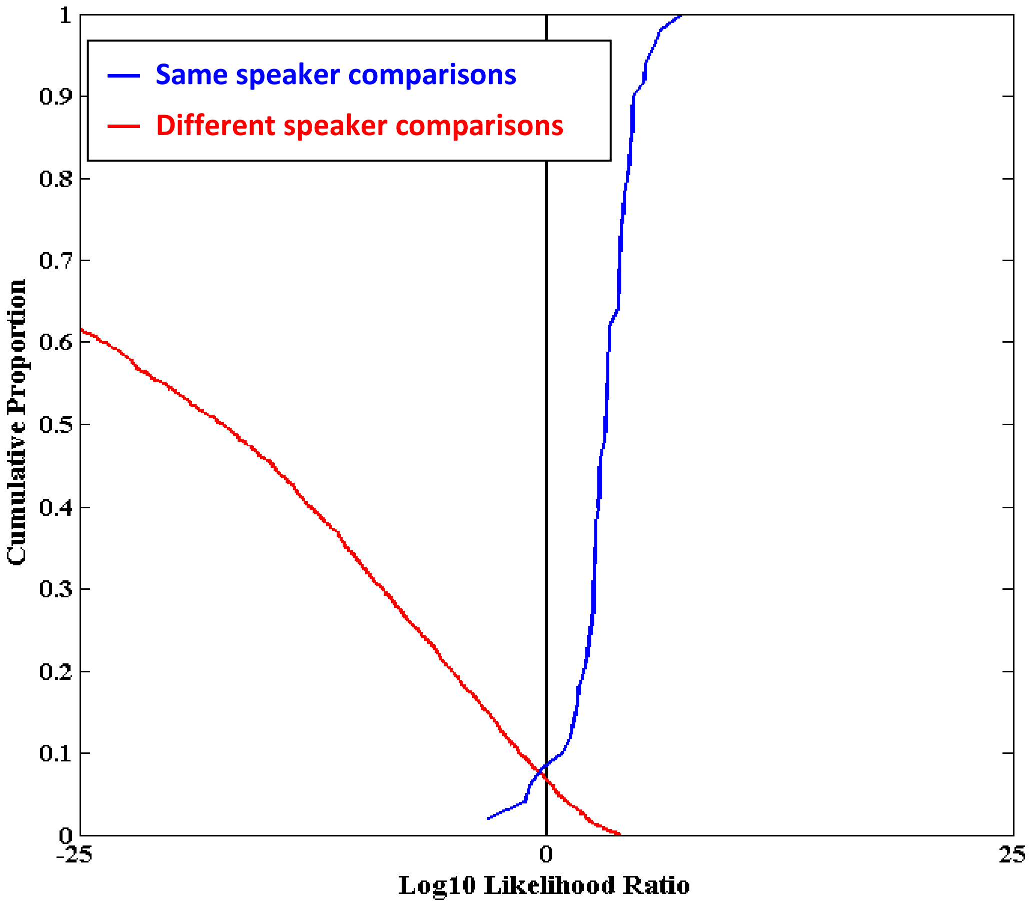


**Figure 8.14:** Zoomed-in Tippett plot of the complete system

Figures 8.13 and 8.14 illustrate the distribution of SS and DS pairs. The strength of evidence of the DS pairs is higher than the strength of evidence offered by the SS pairs. Following Champod and Evett (2000), the system has the potential to offer strength of evidence (either for the prosecution or defense hypotheses) that is considered very strong support. Figure 8.14 shows that the crossover between the curves representing the comparison of SS and DS pairs is very close to the zero threshold, but not on it, and it is possible that calibration of the system might improve its validity (see § 8.5.3 for analysis).

Although the complete system in Table 8.5 includes all available parameters it is necessary to consider the possible performance of other combined systems should they outperform the complete system. Table 8.6 provides ten alternative systems to the complete system from Table 8.5. The organization of Table 8.6 follows that of Table 8.5.

**Table 8.6:** Summary of LR-based discrimination for alternative systems (100 speakers)

| Comparison | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **LTFD1+LTFD2-4+F0 SS** | 92.00 | 5.691 | -3.517 | 7.348 | .0631 | .4322 |
| **LTFD1+LTFD2-4+F0 DS** | 92.82 | 1.528 | -infinity | 4.151 | | |
| **LTFD1+LTFD2-4+AR SS** | 98.00 | 3.811 | -1.423 | 5.380 | .1310 | .6101 |
| **LTFD1+LTFD2-4+AR DS** | 73.06 | 1.311 | -infinity | 4.082 | | |
| **LTFD1+LTFD2-4 SS** | 94.00 | 3.807 | -1.618 | 5.426 | .1361 | .6348 |
| **LTFD1+LTFD2-4 DS** | 71.43 | 1.387 | -infinity | 4.508 | | |
| **LTFD1+F0+AR SS** | 86.00 | 2.432 | -3.900 | 3.594 | .0709 | .4780 |
| **LTFD1+F0+AR DS** | 95.43 | 0.134 | -infinity | 2.373 | | |
| **LTFD1+F0 SS** | 82.00 | 2.150 | -4.335 | 3.353 | .0771 | .5266 |
| **LTFD1+F0 DS** | 94.41 | 0.096 | -infinity | 2.306 | | |
| **LTFD1+AR SS** | 76.00 | 1.046 | -1.967 | 2.143 | .2284 | .7873 |
| **LTFD1+AR DS** | 77.14 | 0.083 | -infinity | 2.101 | | |
| **LTFD2-4+F0+AR SS** | 96.00 | 5.220 | -2.149 | 6.887 | .0647 | .4160 |
| **LTFD2-4+F0+AR DS** | 89.22 | 1.469 | -infinity | 3.848 | | |
| **LTFD2-4+F0 SS** | 96.00 | 5.249 | -2.585 | 6.933 | .0707 | .4742 |
| **LTFD2-4+F0 DS** | 88.20 | 1.465 | -infinity | 3.817 | | |
| **LTFD2-4+AR SS** | 100.00 | 3.319 | 0.457 | 4.951 | .0929 | .8413 |
| **LTFD2-4+AR DS** | 60.20 | 0.789 | -infinity | 3.213 | | |
| **F0+AR SS** | 88.00 | 1.625 | -2.959 | 2.457 | .0855 | .4197 |
| **F0+AR DS** | 91.47 | 0.048 | -268.938 | 2.143 | | |

None of the alternative systems in Table 8.6 outperforms the complete system in terms of validity. The next best performing system in terms of EER (after the complete system) is that of LTFD1+LTFD2-4+F0 with an EER of 0.0631 and Cllr of 0.4322. This second-best system is identical to the complete system minus the inclusion of AR, which suggests that the inclusion of more parameters improves the system’s validity.

### 8.5.3 Overall Likelihood Ratio Results: Calibrated

Calibration is a procedure for improving a system's precision, whereby a well-calibrated system is considered to be more reliable (DeGroot and Fienberg, 1983). Calibration was first utilized by weather forecasters (DeGroot and Fienberg, 1983), but has since made its way into automatic speaker comparison (Ramos-Castro et al., 2006), and phonetic/linguistic-based FSCs (Morrison, 2012). Ramos-Castro et al. (2006, p. 6) have shown the importance of the calibration of LR values computed by an automatic system, arguing that "highly discriminant likelihood ratios might achieve a high performance in terms of probability of error of the posterior probabilities. However, a high calibration loss[46] in the computed LR values may lead to arbitrarily high errors." For this reason, logistic-regression calibration (using a cross-validation method) has been applied here to the complete system in two different orders to compare calibrated results. Figures 8.15 and 8.16 illustrate the first method, in which parameters were calibrated individually and then combined. Figure 8.17 illustrates the results of the second method, where individual parameters were combined and the complete system was then calibrated.

[46] Quantified according to the degree to which LR values incorrectly support a hypothesis.

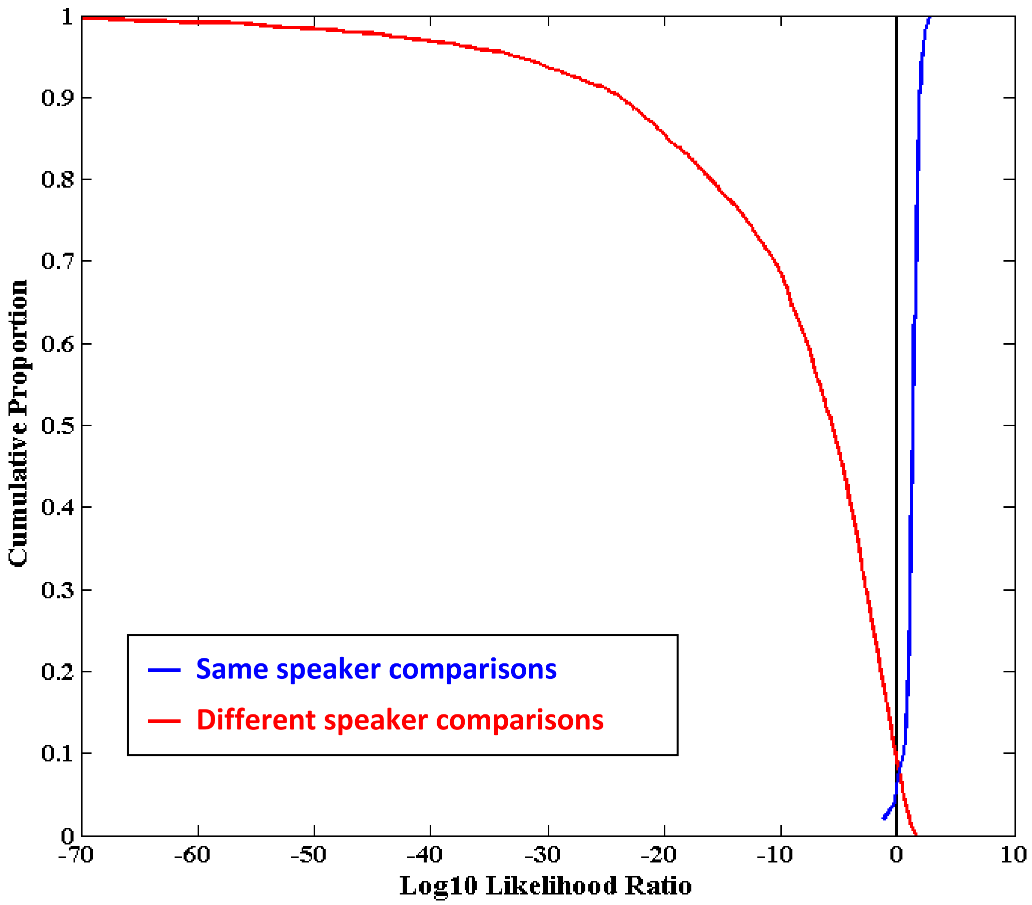


**Figure 8.15:** Tippett plot of the complete system - parameters calibrated individually and then combined

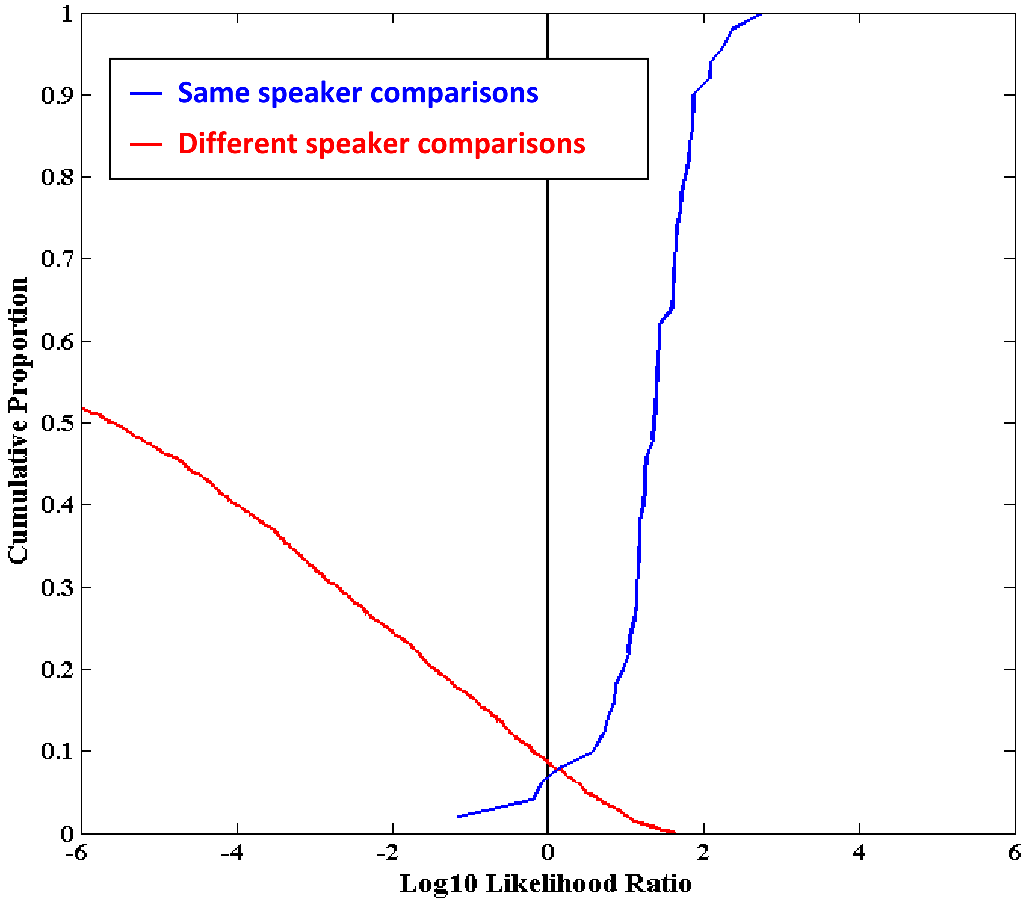


**Figure 8.16:** Zoomed-in Tippett plot of the complete system - parameters calibrated individually and then combined (-6 to 6 LLR)

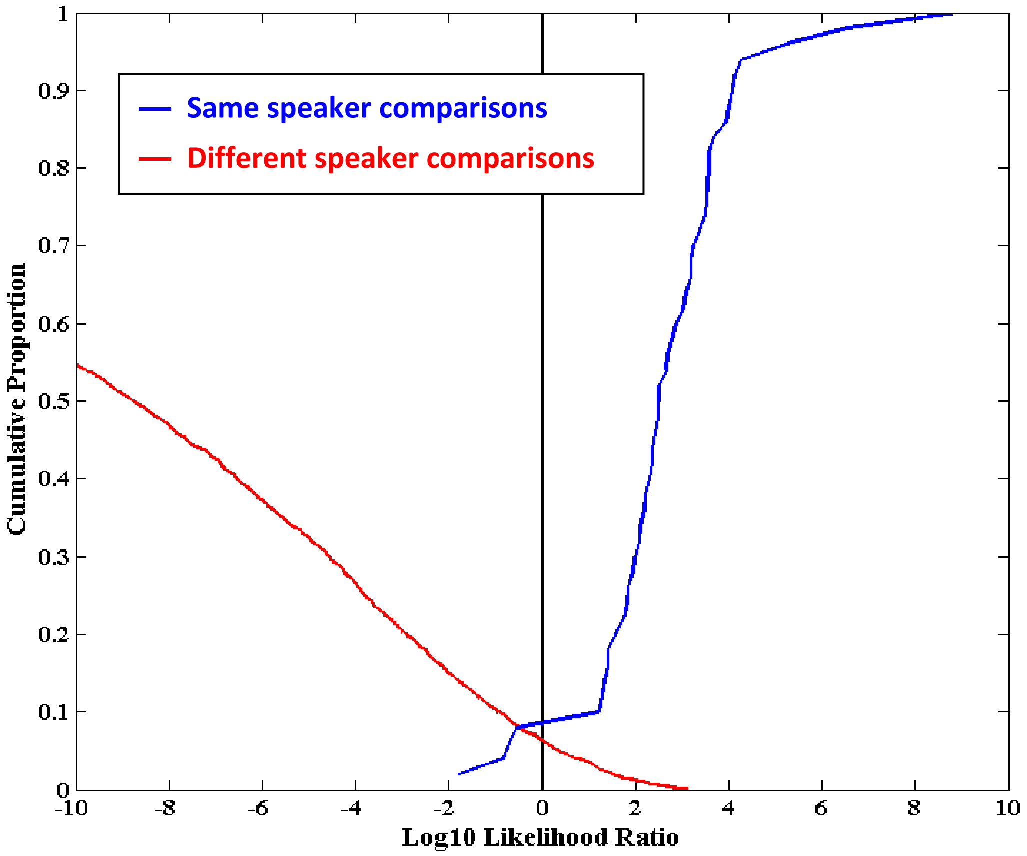


**Figure 8.17:** Zoomed-in Tippett plot of the complete system- system calibration after combination of parameters (-10 to 10 LLR)

The calibration of individual parameters before combination (see Figures 8.15 and 8.16) resulted in an EER of .0554 and a Cllr of .2831. There was an improvement in both the EER and Cllr from the uncalibrated system of .0053 and .0962, respectively. The calibration of the complete system after the combination of parameters in Figure 8.17 resulted in an increase (i.e. a higher value) of EER of .0011, and an improvement (i.e. a lower value) of Cllr of .1408. Results show that for the complete system, calibration before combination provides the best EER, while calibration after combination provides the best Cllr. The differences between the two methods are minimal. However, one improves the gradient result for incorrect/correct judgments (Cllr) while the other improves the hard detection error rate (EER). In forensic speaker comparison, a protocol for the order in which the application of calibration should take place has not been previously discussed. Therefore, more research is needed on the effects of the order of operations in which calibration is applied. It is also entirely possible that calibration could be applied twice, once before combination and once after. However, that was not tested here.

## 8.6 Discussion

This section focuses on three key discussion points: whether clicks can help improve the performance of the complete system, the comparison of results from all systems combinde to the results found for the individual parameters, and potential limitations of a combined system.

### 8.6.1 Do Clicks Improve the Complete System?

The uncalibrated complete system achieved an EER of .0607, while correctly identifying 92% of SS pairs and 93.27% DS pairs. The complete system, however, did not include click rate as one of the combined parameters, as click rate did not lend itself to the calculation of numerical likelihood ratios. The system should now consider whether click rate has the potential to help in discriminating between the SS and DS pairs that were judged incorrectly.

A total of 165 DS pairs (out of a possible 2450 DS pairs) were judged incorrectly by the complete system. Four SS pairs were judged incorrectly. When those 169 incorrectly-judged pairs are extracted, it is possible to identify those pairs that include any of the speakers that had extreme outlying click rates (see § 7.5.3). The three extreme outliers in terms of click rate (speakers 007, 024, and 033) form one half of the pairings in 20 DS pairs. If those extreme click rates were considered within the complete system, one could propose that those 20 DS pairs would then be judged correctly. This would then increase the percentage of correct DS pairs to 94.08% (an additional increase of 0.81%). The extreme click rate outliers are not a part of any of the incorrectly-judged SS pairs, so click rate will not help further the comparison in these pairings.

The small increase in correctly-identified DS pairs would be unlikely to improve the EER or Cllr dramatically. However, it is possible that it would improve the system performance to some degree. Although click rate could be used to help discriminate those SS and DS pairs that were incorrectly identified, there is the potential for click rate to also decrease the system performance. If click rate were to make the correctly-identified SS pairs significantly more dissimilar and the correctly-identified DS pairs significantly more similar, the performance of the complete system would be decreased further. However, it is important to note that this discussion of system performance (where click rate is included) remains hypothetical without including all of the click rate data. As we saw in Chapter 7, click rate appears to be highly variable within and between speakers, which in turn characterizes click rate as an unstable parameter (to a higher extent than AR, even). The inclusion of this unstable parameter could cause more variation in OLRs, which could in turn weaken the system. It is also important to consider that Aitken and Gold (2013) showed that the LRs produced for click rate were associated with relatively weak strength of evidence. Therefore, perhaps the inclusion of click rate in all 2500 SS and DS comparisons may not contribute significantly to the OLRs, leaving the performance of the complete system relatively unchanged.

### 8.6.2 Comparing Individual Parameters to the Systems

The uncalibrated complete system in § 8.6.2 performed relatively well in respect of the system's validity, with a number of the alternative systems' performances following closely behind. The extent of their achievements is best shown in juxtaposition with the performance of the individual parameters prior to their being placed into a combined system. Table 8.7 contains all the individual parameters' performances. The organization of Table 8.7 is identical to that of Table 8.5 and 8.6.

**Table 8.7:** Summary of LR-based discrimination for individual parameters (100 speakers)

| Comparison | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **LTFD1 SS** | 72.00 | 0.224 | -2.158 | 1.902 | .2806 | .8840 |
| **LTFD1 DS** | 71.70 | -4.858 | -68.768 | 1.993 | | |
| **LTFD234 SS** | 74.00 | 0.649 | -8.461 | 4.996 | .0798 | .9023 |
| **LTFD234 DS** | 95.31 | -25.812 | -166.915 | 3.046 | | |
| **F0 SS** | 92.00 | 0.958 | -3.404 | 1.936 | .0849 | .4547 |
| **F0 DS** | 89.90 | -24.204 | -269.159 | 1.906 | | |
| **AR SS** | 90.00 | 0.180 | -1.480 | 2.060 | .3340 | .8981 |
| **AR DS** | 46.20 | -2.940 | -8.760 | 0.820 | | |

Table 8.7 shows that LTFD234 has the lowest EER at .0798, followed by F0 at .0849. LTFD1 and AR both have EERs around .30. The best-performing individual parameter in terms of Cllr is F0 at .4547, with the remaining three parameters close to .90. The results for the individual parameters would suggest that a system including LTFD234 and F0 will have the best opportunity of performing well in terms of system validity. It also appears that a system that includes AR will benefit from SS comparisons, but will potentially be weakened by its DS comparisons.

Table 8.8 provides a summary of the improvements and deteriorations in the complete system performance compared to the best-performing individual parameter for a given LR statistic. The first column identifies the LR statistic, and the second column indicates whether or not the complete system improved, deteriorated, or stayed the same in comparison to the best-performing individual parameter (i.e. LTFD1-4, AR, or F0). The final column indicates the degree of the change identified in the second column.

**Table 8.8:** Performance comparison between individual parameters and the complete system

| Results | Change | Degree of Change |
|---|---|---|
| EER | Improve | .0191 |
| Cllr | Improve | .0754 |
| SS % Correct | Same | 0.00 |
| DS % Correct | Deteriorate | 2.04% |
| SS Mean LLR | Improve | 4.715 |
| DS Mean LLR | Deteriorate | 27.380 |
| SS Min LLR | Deteriorate | 1.602 |
| DS Min LLR | Improve | Infinity |
| SS Max LLR | Improve | 2.32 |
| DS Max LLR | Deteriorate | 3.143 |

Table 8.8 shows that the complete system outperformed any individual parameter in terms of EER, Cllr, SS Mean LLR, DS Min LLR, and SS Max LLR (highlighted in light blue). The complete system deteriorated in performance with respect to DS % Correct, DS Mean LLR, SS Min LLR, and DS Max LLR (highlighted in dark blue). There was no change observed between the complete system and best-performing individual parameter in the performance of SS % Correct (highlighted in mid-blue). Overall, Table 8.8 shows that the complete system has (most importantly) the best system validity as well as the most improvement in terms of strength of evidence for SS. The strength of

evidence for a single parameters tends to be very limited for SS (see Table 8.7), despite the DS strength of evidence being relatively strong. Therefore an improvement in strength of evidence for SS comparisons is highly desired.

The same LR statistics presented in Table 8.8 are also considered with respect to the best overall performing system or individual parameter in Table 8.9. Table 8.9 presents the LR statistic in the first column and identifies which system or individual parameter performed the best in that respect.

**Table 8.9:** Best-performing system or individual parameter in relation to LR statistics

| Results | Best System/Individual Parameter |
|---|---|
| EER | Complete System |
| Cllr | Complete System |
| SS % Correct | LTFD234+AR System |
| DS % Correct | LTFD1+F0+AR System |
| SS Mean | LTFD1+LTFD234+F0 System |
| DS Mean | LTFD234 |
| SS Min | LTFD234+AR System |
| DS Min | Complete System |
| SS Max | LTFD1+LTFD234+F0 System |
| DS Max | AR |

Table 8.9 identifies whether the complete system (light blue), an alternative system (mid-blue), or an individual parameter (dark blue) performed best for the given LR statistic. The complete system remains the best in terms of system validity (which is the most important of the statistics). The alternative systems achieve the best performance for five of the LR statistics, while individual parameters are the best performing for two of the LR statistics. The results in Table 8.9 confirm the opinions set out by experts in § 3.10 that the inclusion of more parameters results in better overall speaker discrimination (i.e. EER). However, the level at which ‘more is better’ is not all-encompassing with

respect to all the LR statistics. It is the case that there are smaller systems and two individual parameters that outperform the complete system in some respects. Therefore, the general opinion of experts - that more parameters are better for discriminating between speakers - needs to be redefined insofar as it does not appear to be the case that the more parameters that are used for speaker discrimination, the better the system validity.

### 8.6.3 Limitations

Despite the relatively good performance of the complete system in terms of validity it still has four important limitations to consider. The first limitation is that the complete system is not outperforming alternative systems and individual parameters in relation to all of the LR statistics. As discussed in § 8.7.2, the inclusion of more parameters does not necessarily correspond to better performance in all respects. Results suggest that the addition of the 'right' parameters could increase performance, but the addition of poorly-performing individual parameters may not improve the overall system. This poses a dilemma regarding which parameters to include in the FSC analysis. Should the expert only select the best-performing parameters, in terms of EER, or should he/she try to include all parameters that characterize an individual's speech? Additionally, the expert must recognize that with the addition of parameters comes the (potential) additional uncertainty introduced in the system. This issue also needs to be addressed in respect of some type of confidence interval, and it may be the case that the confidence interval's measure of credibility will be what sets the complete system apart from an alternative system or individual parameter.

The second limitation to the complete system lies in the steps taken before the combination of parameters, in which dependencies are tested between and within parameters. In general, pairs of parameters that had correlation coefficients of less than 0.25 were considered to be independent of one another. It is possible that the complete system is limited in respect of the combination of parameters that exhibit some small levels of correlation that go unaccounted for. An example is the naïve LR calculation approach taken in Chapter 5 for LTFD1-4. MVKD was used to calculate LRs for LTFD1-4. However, the current chapter found LTFD1 to be independent of LTFD234. This would dictate treating LTFD1 separately from LTFD234 (as was seen in § 8.6.2), and the LRs from the two sets to be multiplied following naïve Bayes. The EER for the MVKD combination was 0.0414, while the current chapter reports an EER for LTFD1+LTFD234 of 0.1361. The two methods for combining LTFD1-4 lead to a dramatic difference in EER. It appears that perhaps not taking small correlations into account when working with the given data has caused EER to increase.

The third limitation is the application of calibration. There is no set protocol for when calibration is to be applied. In § 8.6.3 it was demonstrated that different results can be achieved when the individual parameters are calibrated separately and then combined (yielding an EER of 0.0618), compared to the combination of parameters followed by the application of calibration (giving an EER of 0.0554). One could plausibly consider the calibration of parameters separately before combination, and the calibration of the system in a second phase after the combination of parameters has been carried out. More research is needed in order to make an educated decision on the order in which combination and calibration are performed.

The final limitation is very basic, but it is perhaps the most important. It concerns the threshold at which two parameters are deemed to be dependent on one another. A threshold of 0.25 was selected, but this was somewhat arbitrary. A better understanding is needed through a combination of empirical testing and theory to allow for more reliable decisions to be made on the (in)dependence of parameters. To some extent this is being explored in the International Association of Forensic Phonetics and Acoustics-funded grant entitled 'Identifying correlations between speech parameters for forensic speaker comparisons' (Gold and Hughes, 2013). However, further research is still needed on the relationships between speech parameters in other accents and languages.

## 8.7 Conclusion

The results of this study have shown that the combination of parameters into a complete system improves system performance in terms of validity (EER and Cllr). It is not necessarily the case that more parameters will improve all aspects of the system, but where it matters most - in terms of validity - the addition of more parameters prevails. The combination of the parameters central to this thesis (AR, LTFD, F0, and to some extent clicks) raises the question of what will happen when other parameters are added to the system. Following expert opinion in this respect (see Chapter 3), one would expect validity to further improve. However, it could be the case that there will be a threshold at which the addition of parameters can no longer improve validity. It is also possible that the addition of unstable (highly variable) parameters will make the performance of the system deteriorate. Extrapolating from the current results it is expected that the strength of evidence for SS pairs will only increase as parameters are added, while the strength of evidence will remain similar for DS pairs (as it is already very strong).

It is difficult to predict the performance of a system that includes additional parameters that may exhibit different variation characteristics from the current parameters. It is also difficult to extrapolate the performance of the system while considering parameters (here, click rate) which cannot be incorporated into a numerical LR. For this reason, the current complete system can only serve as a building block contributing towards a larger system that will incorporate a much wider range of linguistic and phonetic parameters, and which will possibly improve the discrimination level of speakers in FSCs.

# Chapter 9: Discussion

In this chapter, the results from the previous six chapters (3-8) are summarized and discussed. The results of the first international survey of FSC practices are evaluated with respect to four of the valued phonetic and linguistic parameters selected by expert forensic phoneticians. The four parameters are then considered individually with regards to their speaker discriminant ability, strength of evidence, and validity. Finally, the combination of the four parameters into a human-based speaker comparison system is discussed and compared with those used in ASR analysis.

## 9.1 Summary of the Forensic Speaker Comparison Practices Survey

The results of the first international survey of forensic speaker comparison practices showed a fundamental lack of consensus on the methods employed in FSCs. Although the finding might come as a surprise to phoneticians and linguists working outside FSS, the degree of variation in methods will not be surprising to those working in various other fields of forensic science (see the journals *Science and Justice* or *Forensic Science International* for a plethora of articles debating forensic methodologies). Most importantly, the survey gave an insight into which parameters experts identified as being the most helpful speaker discriminant parameters above all others. The following are the top five ranked parameters (in order):

1. Voice quality
2. Dialect/ accent variants and *vowel formants*
3. Speaking tempo and fundamental frequency

4. Rhythm
5. Lexical/grammatical choices, vowel and consonant realizations, phonological processes, and fluency

The majority of experts indicated that despite some individual parameters being good speaker discriminants, it is the combination of parameters that hold the most discriminant power in FSCs. The discriminant potential of speech parameters in combination rather than on their own is not often addressed in the research literature.

## 9.2 Summary of Phonetic/Linguistic and Forensic Findings for Individual Parameters

Four parameters identified by survey participants as having a high discriminant value were investigated, namely articulation rate (AR), long-term formant frequencies (LTFD), fundamental frequency (F0), and clicks. This involved assessments of the individual parameters as speaker discriminants (percentage of correctly-classified SS and DS pairs, strength of evidence, and validity), how well expert expectations of the parameters matched the results, and whether the results were similar to those reported in previous studies.

### 9.2.1 Articulation Rate

Speaking tempo, and particularly AR, was identified by 20% of experts in Chapter 3 as one of the most helpful speaker discriminants (ranked 3rd in § 9.1). The high expectations surrounding the discriminant capacity of AR motivated empirical testing. Three key observations can be made in relation to the influence methodology has on the calculation of AR: (1) the definition of the speech interval does not significantly affect results, (2) varying the minimum number of syllables in a speech interval does not make AR significantly more stable, and (3) testing suggests that the exclusion of speech segments, and perhaps the definition of the syllable (i.e. phonetic versus phonological) may have more effect on ARs than other factors.

The findings tell a different story from that predicted by expert opinion, suggesting that there are misconceptions about the discriminant capacity of AR. Under an LR framework, SS pairs were correctly identified 90% of the time, while DS pairs were correctly identified at a rate less than chance (46.2%). Articulation rate contributed weak strength of evidence for SS pairs, and only moderate strength of evidence for DS pairs. AR had an EER of 0.3340 (the highest EER of the three parameters tested under the LR framework, i.e. AR, LTFD, and F0) and a Cllr of 0.8981. AR as an individual speaker discriminant was found to be rather weak. A simple impressionistic determination of speaking tempo, rather than a tedious and potentially unnecessary quantitative analysis of AR, may be sufficient in most forensic cases. Despite apparent misconceptions about the discriminant power of AR, it should nevertheless remain a tool in a forensic phonetician's toolbox as there will always be the possibility of outlying speakers for which AR may be extremely valuable.

### 9.2.2 Long-Term Formant Distributions

Vowel formants, including long-term formant distributions (LTFD), were identified by 28% of experts (Chapter 3) as being among the most useful speaker discriminants (ranked 2nd in § 9.1). This provided the motivation for further discriminant testing. The results from the analysis of LTFD provide both phonetically- and forensically-relevant results. In terms of the phonetic findings, there are two pertinent observations in relation to methodology and speaker specificity: (1) small changes in the package length for LTFD have only a small effect on results, and (2) higher formants (LTFD3 and LTFD4) are suggested to carry a greater amount of speaker-specific information than lower ones.

The forensic findings confirm experts' expectations regarding the discriminant potential of vowel formants. Under an LR framework, the combination of LTFD1-4 correctly identified SS pairs 84% of the time, while DS pairs were correctly identified 97.4% of the time. As a system, LTFD1-4 had an EER of 0.0414 (the lowest EER of the three parameters tested under the LR framework: AR, LTFD, and F0) and a Cllr of 0.5411. Despite the promising findings of LTFD1-4 as a combined system, § 8.4.2 unexpectedly revealed that LTFD1 was statistically independent of LTFD2-4, and should technically be treated separately (as an independent parameter). If LTFD1 is treated separately, the combined LTFD2-4 system still achieves a low EER of .0798 and a Cllr of 0.9023, where SS pairs and DS pairs are correctly identified 74% and 95.3% of the time, respectively. These findings, in combination with previous findings from Becker at al. (2008), Moos (2010), French et al. (2012), and Jessen et al. (2013), suggest that LTFDs perform very similarly to MFCCs under comparable data conditions, and, as an individual speaker discriminant, LTFD is rather strong. The only potential limitation of LTFD is that it averages across all vowels, which in turn eliminates idiosyncrasies and habituations of certain vowels that relate accent information. Unless a single vowel phoneme can yield more promising results, the evidence suggests that LTFD should be considered over individual vowel analysis under the LR framework[47].

[47] Including both could be seen as doubling evidence, insofar as LTFD measurements encompass the multiple formant measurements made for individual vowel phonemes.

### 9.2.3 Long-Term Fundamental Frequency

Fundamental frequency (F0) was identified by 20% of experts (Chapter 3) as being one of the most helpful speaker discriminants (ranked 3rd in § 9.1). This motivated empirical testing of the parameter. The results from the analysis of F0 provide both phonetically- and forensically-relevant results. In terms of the phonetic findings, there are two main observations. Firstly, small changes to the package length of F0 only have a small effect on the results (as we saw with LTFD). Secondly, it was reported in § 8.4.3 that, unexpectedly, F0 did not correlate with LTFD1-4. Given previous research (Narang et al., 2012; Syrdal and Steele, 1985) it might have been expected that F0 and LTFD1 would be correlated, especially since using Lombard speech it has been shown that F1 increases as F0 increases (Kirchhübel, 2010). The independence of F0 and LTFD1 was also reported by Moos (2010), and may be an indication that F0 and F1 correlations can only be found when vowels are analyzed individually as phonemes (Narang et al., 2012; Syrdal and Steele, 1985). However, once F0 is compared to an LTFD that relationship is lost, perhaps because (i) F0 and F1 are not correlated for all phonemes (and an averaging of phonemes eliminates any strong correlation present in the data), or more likely (ii) there is non-vowel information included in the acoustic signal which suppresses any correlation that might be present.

The forensic findings confirm experts' general expectations regarding the discriminant power of F0. Under an LR framework, F0 correctly identified SS pairs 92% of the time, and DS pairs 89.9% of the time. F0 contributed a rather weak strength of evidence for SS pairs, while DS pairs had a much stronger strength of evidence. As a system, F0 had an EER of 0.0849 (the second highest of the three parameters tested under the LR framework: AR, LTFD, and F0) and a Cllr of 0.4547. Given the findings and the plethora of previous LR research on F0, it is suggested that F0 as an individual speaker discriminant is rather strong within a contemporaneous recording. However, it is not entirely clear how well F0 can discriminate between individuals when same-speaker evidence comes from non-contemporaneous recordings. Previous literature would suggest that F0's discriminant power will decrease when same-speaker evidence from different recordings is introduced in addition to any deletrious (external) factors (e.g. disguise, recording transmission, vocal effort; see § 2.2 for more factors). The study by Boss (1996) gave an example of F0 mismatch in a real forensic case. The difference between the F0 in the criminal and suspect recording was 88Hz, due in large part to situational differences in the recordings (the suspect sounded more nervous in the criminal recording than the suspect recording (Boss, 1996, p. 156)). Unless it were to transpire that F0 is robust to many of the factors detailed in § 2.2, the mere comparison of mean F0s and SDs is on its own unlikely to advance the speaker comparison task dramatically. However, as always exceptions are to be made for those individuals who can be classed as outliers, and using F0 in conjunction with other speech parameters for FSCs is suggested.

### 9.2.4 Click Rate

Non-linguistic parameters (which include clicks) were identified by 18% of experts (Chapter 3) as being amongst the most useful speaker discriminants. Again, this motivated empirical testing. The results from the analysis of click frequency provide both phonetically- and forensically-relevant results. In terms of the phonetic findings, there are two pertinent observations: (1) discourse analysis classifications can lend themselves to the quantification and categorization of speech parameters (Wright, 2005), and (2) accommodation effects are present in clicks, in that the rate of clicking, rather than being solely a property of an individual's speech production practices, might arise from an interaction between speaker and interlocutor.

The forensic findings suggest that there are misconceptions surrounding the discriminant capacity of clicks. While it would be dangerous to generalize beyond the variety of English analyzed in this thesis, the view of those forensic practitioners surveyed in Chapter 3 who considered tongue clicking to be a highly discriminant feature of speaker behavior is largely unsupported by the present data for young male speakers of SSBE. Click data is positively skewed and discrete, and there is currently no method available for deriving an LR from them (although see Aitken and Gold (2013) for current developments in proposed algorithms for calculating the LRs of clicks). For this reason, the discriminant capacity can only be assessed qualitatively and with reference to the population statistics. Further, there is insufficient variation across the majority of speakers analyzed for the variable to provide a reliable index of speaker individuality. Additionally, even for the high-rate clickers who stand apart from the majority, there is within-conversation instability to the extent that one would need speech samples of a length seldom encountered in criminal forensic recordings in order to reliably establish an overall click rate. Given the high degree of intra-speaker variation and restricted inter-speaker variation, clicking frequency is a rather weak parameter. Unless it were to transpire that patterns of clicking behavior are different for other varieties of English or differ in accordance with speaker age or gender - and nothing has been found in the sociolinguistic literature on English to support that view - the mere comparison of click rates across samples is, in the overwhelming majority of cases, unlikely to advance the speaker comparison task. However, as with AR, it is advised that clicks remain a tool in an expert's toolbox for those speakers identified as outliers.

## 9.3 Summary of Discrimination Performance by the Overall System

This section has outlined the main forensic findings for articulation rate (AR), long term formant frequencies (LTFD), fundamental frequency (F0), and clicks as a combined, overall system. The forensic findings are assessed with respect to the overall system's success at discriminating between speakers (percentage of SS and DS pairs correct, strength of evidence, and validity), how well expert expectations of the parameters corresponded with the results, and if the results are similar to those found in previous studies.

A large majority of the experts discussed in Chapter 3 indicated that it is the combination of speech parameters that makes for better performance at speaker discrimination than individual parameters. Therefore, the combination of the parameters in this thesis was motivated by expert opinion and the lack of human-based systems that test speaker discrimination of parameters in combination. Under an LR framework, the complete system (LTFD1, LTFD2-4, AR, F0) correctly identified SS pairs 92% of the time, and DS pairs 93.3% of the time. The combined system contributed very good strength of evidence for SS pairs, and even stronger strength of evidence for DS pairs. Overall, the combined system had an EER of 0.0607 and a Cllr of 0.3793. After the complete system was calibrated, the EER and Cllr decreased to 0.0554 and 0.2831, respectively. The combination of parameters into a complete system therefore improves system performance (in comparison to individual parameters) in terms of validity (EER and Cllr). It is not necessarily the case that the inclusion of more parameters (e.g. SS % correct, DS % correct, mean SS LLR) improves all aspects of the system, but where it matters most (validity) the addition of parameters does result in an improvement. The results of the complete system are almost as good as those from ASRs under similar conditions (French et al., 2012). Table 9.1 compares the results of the research presented in this thesis against those from the ASR in French et al. (2012).

**Table 9.1:** Human-based results against ASR (Batvox) results from French et al. (2012) on studio quality data

| | **Same Speaker** | **Different Speaker** |
|---|---|---|
| **Current study** | 92% | 94.1% |
| **French et al. (2012)** | 100% | 95% |

The percentages presented in Table 9.1 indicate the proportion of SS and DS comparisons judged correctly where studio-quality recordings were used. Given that the system developed in this thesis only incorporates three parameters under an LR framework, the incorporation of more speech parameters might improve system performance further.

Out of all the views advanced by experts that were reported in Chapter 3, perhaps the most valuable and the most accurate is that the combination of parameters results in the best speaker discrimination performances. The current system serves as a starting point from which to expand. Additional research needs to be carried out on the discriminant power of parameters in combination.

## 9.4 Overall Findings

There are two principal findings that emerge from the present research: (1) the performance of the human-based system created in this thesis is comparable to ASR performances on the same studio-quality recordings, and (2) the FSS community is faced with many obstacles if they wish to continue to align themselves with other, more developed, forensic sciences by implementing an LR framework for FSCs.

### 9.4.1 Human-Based System versus ASR

The research conducted for this thesis was not intended to provide a comparison of the efficacy of human-based systems and their ASR counterparts. Nevertheless, through the development of this human-based system over the course of the current research project, the methodologies employed have made quantitative comparisons between the human-based system and ASRs possible. ASRs used for FSCs are typically known for demonstrating their validity through error rates, and the testing of these ASRs is easily replicable. ASR error rates are typically presented in terms of the frequencies of false negatives and false positives. A false negative is the classification of a target trial as a non-target trial, and a false positive is a non-target trial classified as a target trial (van Leeuwen and Brümmer, 2007). The human-based methodology (using phonetic and linguistic parameters) for FSCs has become known for *not* providing error rates and for a lack of replicability. However, it is shown that by adopting a numerical LR framework, a human-based system can also provide validated results, while fostering tests that are easily replicable.

The performance of the human-based system, consisting of LTFD, AR, F0, and click rate, is comparable to that of an ASR tested on the same type of data (high-quality studio recordings). The human-based system created for this thesis reported false positive errors (different-speaker comparisons = LR > 1) of 5.9%[48] and false negative errors (same-speaker comparisons = LR < 1) of 8.0%[49]. An ASR system tested on the same data (French and Harrison, 2010) reported false positive errors of 4.5%, and achieved zero false negatives.

It is important to note here that the performance of a human-based system is dependent upon the expertise of the analyst. It is likely that some degree of cross-analyst variation would be observed. For the present human-based system, LTFD and F0 would be least susceptible to inter-analyst variation given that the methodology for extracting data is relatively automatic[50]. AR and click rate calculations are more dependent on the analyst. For example, when calculating AR the analyst must decide which speech to include in an interval and what to ignore, and for clicks the analyst must decide whether s/he is hearing a click rather than a percussive. For this reason, caution should be exercised so as not to overestimate the replicability of results.

#### 9.4.1.1 A Fair Comparison?

The comparison of the human-based system and an ASR investigated by French et al. (2012) showed that the ASR only minimally outperformed the

[48] This includes click rate.

[49] This does not include click rate.

[50] See Jessen and Becker, 2010 (discussed in § 5.2) and Konrat and Jessen, 2013 for variation in LTFD and F0 results when measured by multiple analysts.

human-based system on studio-quality data. However, such a comparison puts the human-based system at a considerable disadvantage. ASRs have consistently been shown to perform well under high-quality data conditions (Campbell, 1997; French and Harrison, 2010; Reynolds, 2002; Reynolds et al., 2000). The ASR, however, is susceptible to degradation in system performance as the data quality gets worse. French and Harrison (2010) tested an ASR on real case data where the outcomes of the cases were known to the authors (i.e. guilty/not guilty verdicts). For the 767 comparisons undertaken, the ASR achieved an EER of 24.2%. When the ASR was tested on the cases in which the system judged the technical quality of the data to be adequate, it accepted 171 of the 767 comparisons and produced an EER of 5.4%. When the ASR was further tested on recordings that the system judged to be only marginally adequate, an EER of 15.1% was returned for the 369 comparisons. The results indicate a steep fall in performance for the ASR when processing less than ideal quality recordings.

A human-based system may be able to offer a better performance than ASRs when testing lower-quality recordings. For example, the 767 real forensic comparisons used for testing with an ASR by French and Harrison, 2010 had previously been analyzed by the authors using a phonetically- and linguistically-based methodology. None of those 767 comparisons resulted in a known miscarriage of justice, and for all comparisons their conclusions were in agreement with those made by the trier of fact.

In real forensic cases where recordings are of poor quality or short duration, in conclusion, a human analyst may be better equipped to extract data and make conclusions than an ASR.

#### 9.4.1.2 Scope for Improvements in the Human-Based System

The human-based system created for this thesis was limited to four parameters owing to the inherent time restrictions of the research. Given the analysis presented in § 8.6 and § 8.7, it is reasonable to assume that the addition of good speaker discriminants would increase the validity of the system. With enough additional parameters it is possible that the human-based system could eventually outperform an ASR on high-quality studio data. The addition of parameters should not be done ad hoc, but should involve phonetically- and linguistically-informed choices. This would involve: (i) selecting parameters that are good speaker discriminants (e.g. voice quality, VOT, or those parameters reported by experts in § 3.9 or § 3.10,), and (ii) not selecting parameters that are significantly correlated with others (but if they are, weighting the correlations appropriately).

Additionally, it is appropriate to consider the integration of this human-based system with ASRs. This could potentially be similar to the Vocalise[51] software package created by Oxford Wave Research Ltd (Vocalise, 2013) that measures MFCCs in addition to traditional phonetic parameters (e.g. F0 and LTFD; see § 10.2 for further discussion of Vocalise). Given the good performance of the ASR on the data tested, and the good performance achieved by the human-based system, a combination of the two could result in an even better overall performance. If integration was to be done it would probably be best to choose between the inclusion of MFCCs or LTFDs, as they effectively analyze the same aspects of a speaker (i.e. vocal tract resonances). LTFD analysis provides the analyst with an indication of the speaker's habitual use of the vowel space,

[51] http://www.oxfordwave research.com/j2/products/vocalise [Accessed 8 August 2013]

which is strongly correlated with the dimensions of the vocal tract (Moos, 2010; French et al., 2012). MFCC analysis reports on approximately the same aspect, as MFCCs are essentially abstract representations of the dimensions of the vocal tract. For this reason, LTFDs and MFCCs are highly correlated (French et al., 2012) and the inclusion of both parameters would result in the doubling of evidence, which in turn could lead to a miscarriage of justice. Therefore, given that MFCCs (marginally) outperform LTFDs (as shown in § 5.2), it is suggested that MFCCs should be selected over LTFDs. An integrated system could then consist of ASR analysis, F0, AR, and click rate, which again is entirely possible if ASR (MFCC) data are found to be independent of the other parameters included in the integrated system. Without testing this integrated system it is difficult to say with certainty that discriminant performance would improve. However, with the exchange of LTFD for MFCC that is likely to be so.

#### 9.4.1.3 The Trade-Off

When comparing the performance of the human-based system with ASRs, it is important to consider a trade-off that occurs when human intervention is involved. The 'human intervention' is the analysis and examination undertaken by an analyst for a given comparison (e.g. through sound file editing, time taken for an analysis). Irrespective of methodology (human or ASR) the level of human intervention typically needs to be increased[52] as the quality of the recordings (or duration) being compared decreases. In turn, the level of human intervention utilized in a FSC creates a

[52] In ASR analysis more time would be required of the analyst for the editing of recordings (cleaning them up, for example, with a bandpass filter), and for human-based analysis more time would be required for the analysts to extract measurements and additional data from the poor quality recordings.

trade-off with the variation that may be present in the FSC results. The more human intervention required, the more cross-analyst variability may be present in results. For example, if an FSC comparison was vital to a criminal case, rather than having an ASR completely reject a case that was of poor recording quality (or offer a conclusion with a very high EER associated with it), a human-based analysis could be attempted in an effort to extract any relevant conclusion.

### 9.4.2 Obstacles Facing the Implementation of an LR Framework

The exercise of creating a human-based system for this thesis has revealed a number of difficulties that surround the LR framework and its application to FSCs. Those shortcomings are as follows:

1. Subjective elements of the methodological process
2. Delimiting the relevant population
3. Availability of population statistics
4. Lack of models available to calculate LRs
5. Appropriate combination of parameters

The LR framework is intended to create a separation between an expert's bias and the facts of the evidence; however, within FSCs the application of the LR framework still has elements of subjectivity. It is possible to alter methodologies (e.g. package lengths of LTFD or F0, as seen in § 5-6) in order to achieve a more desirable strength of evidence. For this reason, it is argued that the LR framework is not completely objective. However, the levels of subjectivity are far less under an LR framework than with other frequentist frameworks.

The present study did not address the debate surrounding the delimitation of the relevant population, as intrinsic LRs were carried out on the known optimal population (i.e. the test and reference sets of speakers came from the same linguistically homogeneous corpus). However, the delimitation of the relevant population will remain a difficulty facing FSCs, as was also argued in French et al. (2010). For this reason, a mutually agreed protocol or methodology would need to be proposed in order for the continuing debate to subside. See Hughes (in progress) for further information and discussion on the issue of the relevant population.

The lack of population statistics, as previously mentioned by French and Harrison (2007) and French et al. (2010), is a very real problem for the implementation of a numerical LR. The time, effort, and resources needed to collect a sufficient quantity of population statistics are almost limitless. As a testament to this, for the present study it took nearly three years to collect population statistics for just four parameters for 100 speakers in a single and very specific population (and, moreover, these data were extracted from a previously-collected database). All things being equal, following a conservative number of possible parameters to analyze in a FSC of around 60 (just for example), the collection of sufficient population statistics for a delimited population would take approximately 45 years[53] to complete by a single person. At that point, given the occurrence of sound change, it would be time to scrap the collected population statistics and start again. Such a feat hardly seems practical. Therefore, in order to continue the development of the LR framework for FSCs, alternatives need to be put in place.

The lack of appropriate models is currently one of the biggest challenges faced by the FSS community, because as it stands only certain parameters can be included in a numerical LR framework. A FSC does not simply consist of a few vowel formant measurements, and phoneticians would argue that it takes a

[53] 45 years = 60 parameters at 4 parameters collected every 3 years

number of parameters beside vowels to play a role in characterizing an individual's speech. For this reason, more linguistically-motivated models are required to enable the incorporation of previously unrepresented phonetic/linguistic parameters in an LR framework (see Aitken and Gold, 2013 for the proposal of a new, linguistically-motivated model).

Finally, further research should focus on the identification and testing of appropriate methods for combining speech evidence. There are currently multiple options for this, but little testing has been done to compare such methods. It is also the case that only three of these methods (i.e. fusion, MVKD, GMM, and naïve Bayes) are ever really implemented in FSC. If the numerical LR framework is to become the way of the future, then research should consider the use of Bayesian Networks for combining speech evidence; these have already been successfully implemented in more developed forensic disciplines (e.g. DNA; Evett et al., 2002).

## 9.5 Methodological Limitations

As with most empirical research, methodological shortcomings are often inevitable, and the research presented in this thesis is no exception. This section outlines and discusses three general methodological limitations: (1) the absence of non-contemporaneous data, (2) the use of intrinsic LR testing, and (3) using only speakers 1-50 for all the same-speaker comparisons when calculating LRs.

The recordings used in the present study were all obtained from single recording sessions, as there were no non-contemporaneous data available for any of the 100 speakers. Non-contemporaneous speech samples have been a frequent topic of discussion in previous research, with an increasing number of studies citing the importance of incorporating recordings that are made several days, weeks, months, or even years apart (see Enzinger and Morrison, 2012; Loakes, 2006; Morrison et al., 2012b; Nolan et al., 2009; Rhodes, 2013). However, I would argue that there are a number of other external factors that have more of an impact on results than the separation of recordings by a mere few days (such as the effects of accommodation reported in § 7.5.5). In any case, the results presented here, despite the fact that they are based on the use of contemporaneous recordings, nevertheless provide a general baseline in relation to the discriminant abilities and practicalities of a human-based system.

The second methodological shortcoming is the absence of extrinsic criminal samples on which to test the human-based system. The LRs calculated in the present study were based on the DyViS data set of recordings of 100 speakers, whereby the first 50 speakers always acted as the test samples and the second set of 50 speakers always acted as the reference samples. This means that the 100 speakers were in some way a part of either the test or reference sample, and that tests were not conducted using outside data sets. As a result, the testing is susceptible to over- or under-estimation of the strength of evidence as proposed in Rose et al. (2006c, p. 329). However, if an additional relevant database existed, extrinsic testing would be possible, as the population statistics for DyViS are now available.

The final limitation is the way the data (i.e. speakers) were divided for calculating LRs. It is probable that assigning different speakers to act as either the suspect/criminal or background population would cause variation in the resulting LRs. For empirical testing purposes, the calculation of intrinsic LRs typically requires a set of speakers to act as both the criminal and the suspect, while an additional set of speakers must act as the reference population. The number of speakers chosen to act as the criminal/suspect or the reference population can vary, and is entirely at the analyst's discretion. For convenience, this thesis divided the data set evenly. As such, the calculation of LRs produced 50 SS comparisons and 2450 DS comparisons. However, it is entirely feasible for the speakers to have been divided in a number of alternative ways, which might have produced different outcomes.

## 9.6 Implications for Forensic Speaker Comparisons

For forensic phoneticians, the population statistics presented in Chapters 4-7 may serve as a helpful tool for casework. It is also suggested that all forensic phoneticians should consider the inclusion of LTFDs in their analysis. The relative ease of extracting LTFDs from speech recordings means that this parameter could be easily utilized and potentially offers a large contribution to FSCs.

As things currently stand, not all speech parameters can be incorporated into a numerical LR. Therefore, the use of a complete[54] numerical LR is impossible. For this reason, experts are faced with a number of decisions, should they choose to continue to develop methodologies in an effort to align with other forensic disciplines. Judge Hodgson from Australia argues that not all types of evidence can be sensibly assigned an LR, and that therefore there is no way of mathematically combining all evidence under a Bayesian framework (Hodgson, 2002). Should the field of FSS agree on this statement, it is worth considering whether all speech evidence can be sensibly assigned a numerical

[54] By "complete", it is meant that all possible analyzable speech parameters are included.

LR. If all speech evidence can be forced into a numerical LR model then an expert's job is done. However, this does not seem a plausible possibility, given the complexity of speech evidence. For this reason, it is my view that experts are left with three options for the future: they must take the first two if they wish to align with other forensic disciplines, and the third, a default option, if they are content with the status quo:

1. Adopt a verbal form of the likelihood ratio framework
2. Present evidence in the form of numerical LRs (for those parameters that readily lend themselves to such a framework) and present the remaining speech parameters using a different conclusion framework.
3. Continue presenting evidence as they do currently.

Although practitioners of forensic phonetics have accepted that the LR is the logically and legally correct framework within which to present evidence, the practicalities of the framework will inevitably contribute to the direction taken by the majority of experts in the future. A large role in the adoption of an LR framework will also be played by regulations under the country of practice and simply practicality issues. All countries and institutions are constrained by the laws, rules, and regulations in which they work. In § 3.11, an expert from China reported that s/he was required to use a binary conclusion framework by his/her government employer. If forensic phoneticians have the luxury of being able to choose their conclusion framework, their decision will come down to a practicality factor. Given the results of this thesis, I believe that the implications for the field of FSCs are such that a complete numerical LR is unrealistic, and that alternatives should be explored (such as 1 and 2 presented above).

# Chapter 10: Summary & Conclusion

The following chapter reviews the empirical results and discussion detailed in Chapters 3-9 by relating them back to the four research questions introduced in Chapter 2. The chapter concludes with suggestions for future work that would expand upon the research aims of this thesis.

## 10.1 Research Questions Revisited

**(1) What phonetic and linguistic parameters do practicing forensic phoneticians (around the world) typically analyze in a FSC case and which parameters do they recommend as being highly discriminant?**

Chapter 3 provided insight into parameters commonly used in FSCs. Those reported parameters were as follows: vowels (formants), consonants (timing aspects, frequencies of energy loci, auditory quality), F0 (mean, median, mode, SD, range, alternative baseline), voice quality, intonation, speech tempo (AR and SR), rhythm, linguistic features (aspects of turn-taking, patterns of code switching, discourse markers (including clicks), telephone opening and closing behaviors, lexico-grammatical usage), and non-linguistic features (filled pauses, audible breathing, laughter, throat clearing).

Out of all the possible parameters analyzed in FSCs, the experts' top ranked parameters (the first three) in terms of the expected discriminant ability were: (1) voice quality, (2) dialect/accent variants and vowel formants, and (3) speaking tempo and F0. Three of these parameters (LTFD, AR, F0) and clicks were chosen for analysis in subsequent chapters.

### (2) If experts are to provide their opinion on the most helpful speaker discriminants, will these 'selected' parameters be good speaker discriminants?

Chapters 4-7 presented the results of the discriminant ability of AR, LTFD, F0, and clicks. LTFD and F0 showed promising results, with EERs of less than .1. AR and clicks were not as good at discriminating between speakers. AR had an EER of .33, and although LRs were not calculated for clicks, the results suggested that click frequency was a very unstable parameter and unlikely to be a useful discriminant for the majority of speakers.

### a. Do experts' expectations surrounding the discriminant value of certain speech parameters match the parameters' empirically tested performance?

The results presented in Chapters 4-7, in combination with the results from the survey in Chapter 3, suggest that disparities exist between expert opinions and empirical findings with regard to AR and clicks. However, expert expectations appear to be accurate for LTFD and, to a lesser extent, F0.

### (3) How well do speech parameters work in combination to discriminate between speakers?

**Table 8.5:** Summary of LR-based discrimination for the complete system (100 speakers)

| Comparison | % Correct | Mean LLR | Min LLR | Max LLR | EER | Cllr |
|---|---|---|---|---|---|---|
| **Complete System SS** | 92.00 | 5.673 | -3.082 | 7.316 | .0607 | .3793 |
| **Complete System DS** | 93.27 | 1.560 | -infinity | 3.963 | | |

Table 8.5 (from Chapter 8) presents a summary of the uncalibrated LR results for LTFD, F0, and AR in combination. The system combining these three

parameters produces a lower EER and Cllr than any individual parameter. The results supported the view expressed by survey participants, i.e. that the combination of parameters is more successful in discriminating between speakers than any of the parameters individually.

### a. What steps need to be taken in order to appropriately combine speech parameters?

The proper combination of speech evidence will vary depending on the given data set. However, for all cases, correlations should be tested in order to establish whether there are dependent relationships amongst speech parameters in order to avoid miscarriages of justice (through the doubling of evidence). For the given data set, dependent relationships were identified amongst LTFD2-4, and MVKD was used to account for the existing correlations (by applying statistical weightings). The remaining parameters were found to be mutually exclusive and the speech evidence was therefore combined using Naïve Bayes.

### b. Are multiple speech parameters in combination always better than individual parameters at discriminating between speakers? Are more parameters better?

Chapter 8 revealed that for the given data set, the addition of more parameters improved system validity. However, this did not improve all of the LR statistics (e.g. exceptions included SS and DS percent correct, SS and DS Mean LLR, and SS and DS Max LLR).

### (4) What are the practical limitations/implications of using the numerical LR framework in FSCs?

Chapter 9 discussed the limitations and implications of using a numerical LR framework that arose during the development of the system created for this thesis. Those limitations include, but are not confined to: subjective elements of the methodological process, difficulties in delimiting the relevant population, the limited availability of population statistics, the lack of models to calculate LRs for complex speech data distributions, and the range of methods used to combine speech evidence.

**a. What recommendations, if any, can be provided by attempting to implement a numerical LR framework?**

Chapter 9 concluded that two fundamental factors – the amount of time needed to collect enough data to create a human-based system consisting of just four parameters, and the inherent complexity of speech evidence - inhibit the implementation of a numerical LR framework. Although it has been done, it is difficult to implement a completely numerical LR framework if one is to do it properly/responsibly. Chapter 9 therefore recommended that practitioners wishing to use a Bayesian framework should consider adopting a verbal LR framework or a combination of a verbal and numerical LR framework, instead of a purely numerical one.

**b. What can a human-based (acoustic-phonetic) system tell the field regarding the ease with which a numerical LR can be computed for FSCs?**

The algorithms with which a human-based system can calculate a numerical LR are the same as those used by an ASR. However, the time needed to collect and analyze the data for the actual LR calculation is much more intensive for a human-based system. Time constraints aside, Chapter 9 suggested that the human-based system was capable of producing results comparable to those of ASRs. It also proposed that with real forensic material (e.g. degraded or shorter criminal recordings), the human-based system could potentially outperform an ASR.

## 10.2 Opportunities for Future Research

Successful application of the numerical LR framework will require a large amount of additional research if the presentation of FSC conclusions in such a framework is to become an everyday reality. The work presented in this thesis would benefit from future investigations in two prominent areas of research: (1) the integration of ASRs and phonetic-linguistic parameters, and (2) research into more transparent and successful ways for combining correlated speech evidence (e.g. fusion).

As discussed in § 9.4.1.2, there is potential for improvements in speaker discrimination through the integration of ASR techniques and phonetic-linguistic parameters. Future research in this area may benefit from the availability of systems such as Vocalise (Vocalise, 2013) which allows for (semi-)automatic analysis and comparison of samples using LTFDs as well as MFCCs.

The second challenge – how to combine correlated speech parameters in a transparent and appropriate manner – could be explored through the use of Bayesian Networks. Other, more developed, forensic disciplines (e.g. DNA analysis) rely on Bayesian Networks (Aitken and Taroni, 2004) as an explicit and logical method for combining evidence (see Evett et al., 2002). In forensics, the use of Bayesian Networks for the derivation of FSC LRs enables any correlations that exist between (or within) parameters to be weighted appropriately. A Bayesian Network therefore avoids over- or under-estimation of the evidence, by accounting for the correlations through a front-end processing technique. An example of a Bayesian Network (using the parameters analyzed in this thesis) is provided in Appendix B.

## 10.3 Conclusion

This thesis set out to explore viability of aligning the field of FSC with other, more developed, forensic sciences that are currently implementing a numerical LR framework. The research has highlighted a number of difficulties that face the FSS community if experts are to continue in their efforts to align themselves with advanced forensic disciplines. At the present time, the application of a completely numerical LR framework is fundamentally impractical, insofar as the numerical LR is unable to incorporate all pieces of speech evidence that "count" (those which are discrete rather than continuous, and which cannot be adequately quantified). In the process of addressing the main aims of this thesis, additional findings were presented, including: the survey of FSC practices, the discriminant capacity of individual parameters (AR, LTFD, F0, clicks) and those parameters in combination. It is hoped that the findings of the thesis will encourage discussion leading towards the solution of problems involved in adopting a numerical LR framework for FSCs. It is also hoped that this research will prompt forensic phoneticians to consider implementing a verbal LR framework (or a combination of verbal and numerical LRs) in order to mitigate the practical limitations associated with completely numerical LRs.

# Appendix A

## Survey Instructions:

Each participant was emailed a unique, secured web-link that directed them to the survey. The survey had general instructions on the first page, which read:

*'Please keep in mind that there are no right or wrong answers. This survey is meant to serve as a tool to gain insight into the general practices in forensic speaker comparisons around the world as well as finding out which features forensic phoneticians identify as useful speaker discriminants.*

*All answers will be kept anonymous and names of participants will never be revealed, so please answer honestly.'*

Before each of the 9 sections of the survey there were additional instructions to remind the participants to generalize to the best of their ability across all cases they had worked on rather than always responding that the given feature was case-dependent. The instructions for the 9 sections read as below, with X representing the number of questions in the section of questions that the instructions preceded, as detailed in the survey content section of this thesis:

*'The following section consists of X questions. Please answer the following questions with regards to all speaker comparison cases you have worked on. I understand that many features and analyses are case dependent, so please do your best to make generalizations where applicable.'*

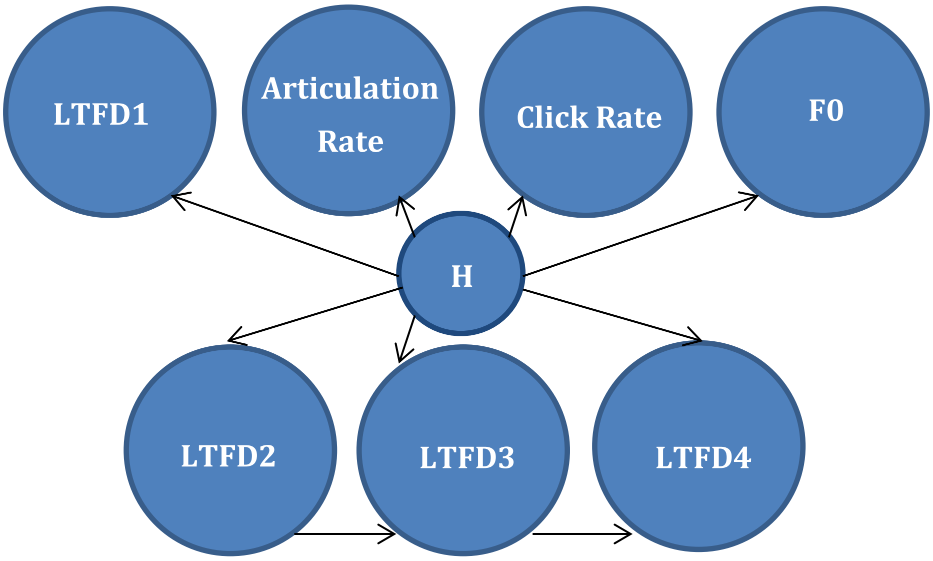
There were some participants who gave the predicted “it depends on the case” answer for questions, but for the most part the majority did an excellent job at generalizing across cases.

# Appendix B

## Example of a Bayesian Network for Speech Evidence:

Bayesian Networks have never been used in FSS before, yet they offer a method for combining evidence that is both transparent and readily accepted by other forensic communities (Evett et al., 2002). For this reason, further research exploring the development of Bayesian Networks for FSC casework would be worthwhile. A hypothetical example (using the parameters analyzed in this thesis) is provided below as an illustrative example of a Bayesian Network using speech evidence, where H represents the hypothesis.

**Figure 11.1: Hypothetical Bayesian Network of speech parameters**



Bayesian Networks, such as the figure above, can be used to calculate Overall LRs (OLRs) provided that probability densities and variances exist for each parameter. If a case does not account for evidence from a certain parameter in the figure above, it is possible to simply leave that node out and the remaining portions of the Network will still be functional for calculating OLRs.

# List of Abbreviations

| | |
|---|---|
| Δ | change or difference |
| AR | articulation rate |
| ASR | automatic speaker recognition system |
| AcPA | acoustic phonetic analysis |
| AuPA | auditory phonetic analysis |
| AuPA + AcPA | auditory phonetic-cum-acoustic phonetic analysis |
| Cllr | cost log likelihood ratio |
| CPS | classical probability scale |
| DS | different speaker |
| DyViS | Dynamic Variability in Speech |
| EER | equal error rate |
| F0 | fundamental frequency |
| FSC | forensic speaker comparison |
| FSS | forensic speech science |
| GMM-UBM | Gaussian mixture model – universal background model |
| HASR | human-assisted automatic speaker recognition |
| IQR | interquartile range |
| LTFD | long-term formant distributions |
| LTFD1-4 | long-term formant distributions one through four |
| LR | likelihood ratio |
| LLR | $\log_{10}$ likelihood ratio |
| LTS | long-term spectrum |

| | |
|---|---|
| MFCC | Mel-frequency cepstral coefficient |
| OLR | Overall likelihood ratio |
| SD | standard deviation |
| SS | same speaker |
| SSBE | Southern Standard British English |
| VOT | voice onset time |
| VQ | voice quality |

# References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Aitken, C. G. G. and Gold, E. (2013). Evidence evaluation for discrete data. *Forensic Science International,* 230(1-3), pp. 147-155.

Aitken, C. G. G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society*, 53 (1), pp. 109-122.

Aitken, C. G. G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd ed.). Chichester: John Wiley & Sons, Ltd.

Alderman, T. (2004). The Bernard data set as a reference distribution for Bayesian likelihood ratio-based forensic speaker identification using formants. *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Macquarie University, Australia., pp. 510-515.

Alwan, A., Narayanan, S. and Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II: the rhotics. *Journal of the Acoustical Society of America*, 101(2), pp. 1078-1089.

Atkinson, N. (2009). *Formant dynamics for SSBE monophthongs in unscripted speech*. Unpublished; University of York. MSc.

Aristotle. *Metaphysica* 10f-1045a.

Arndt, H. J. (1963). Untersuchungen über den Einfluβ der Mandelausschälung auf die menschliche Stimme. *Folia Phoniatrica,* 15, pp. 110-121.

Ash, S. (1988). Speaker identification in sociolinguistics and criminal law. In K. Ferrara, B. Brown, K. Walters & J. Baugh (eds.) *Linguistic Change and Contact.* Austin: University of Texas. pp. 25-33.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contour. *Journal of the Acoustical Society of America,* 52(6) (part2), pp. 1687-1697.

Baldwin, J. R. and French, J. P. (1990). *Forensic Phonetics*. London: Pinter.

Ball, M. J. (1989). *Phonetics for speech pathology*. London: Whurr Publishers Ltd.

Bauer, H. (1963). Die Beeinflussung der weiblichen Stimme durch androgene Hormone. *Folia Phoniatrica,* 15, pp. 264-268.

Bayes, T. and Price, R. (1763). *An Introduction to the Doctrine of Fluxions, and Defence of the Mathematicians against the Objections of the Author of the*

*Analyst.* Printed for J. Noon, London. The Eighteenth Century Research Publications Microfilm A 7173 reel 3774 no. 06.

Becker, T., Jessen, M., and Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models*. Proceedings of Interspeech 2008*. Brisbane, pp. 1505-1508.

Berendes, J. (1962). Veränderungen der weiblichen Stimme durch virilisierende und anabole Hormone. *Folia Phoniatrica,* 14, pp. 265-271.

Bertsch McGrayne, S. (2012). *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*. New Haven and London: Yale University Press.

Bhuta, T., Patrick, L., and Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, 18(3), pp. 299-304.

Boss, D. (1996). F0 and real-life speaker identification. *Forensic Linguistic*s, 3(1), pp. 155-159.

Bouchayer, M. and Cornut, G. (1992). Microsurgical treatment of benign vocal fold lesions: Indications, technique, results. *Folia Phoniatrica,* 44, pp. 155-184.

Braun, A. (1994). The effect of cigarette smoking on vocal parameters. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland: p. 4.

Braun, A. (1995). Fundamental frequency – How speaker-specific is it? In A. Braun and J.P. Köster (eds.) *Studies in Forensic Phonetics*. Trier: Wissenschaftlicher Verlag, pp. 9-23.

Brief Amicus Curiae of Americans for Effective Law Enforcement (1997). Submitted in Kumho Tire v. Carmichael, No. 97-1709, in the Supreme Court of the United States (October Term).

Broeders, A. P. A. (1999). Some observations on the use of probability scales in forensic identification. *International Journal of Speech, Language and the Law,* 6(2), pp. 228-241.

Broeders, A. P. A. (2001). Forensic Speech and Audio Analysis. Forensic Linguistics. 1998 to 2001 A Review, *Proc. 13th INTERPOL Forensic Sciences Symposium*. 16-19 October. Lyon, France.

Brümmer, N., Burget, L., Cernocký, J. H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., and Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the

NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing,* 15, pp. 2072-2084.

Brümmer, N. and du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language,* 20(2-3), pp. 230-275.

Byrne, C. and Foulkes, P. (2004). The mobile phone effect on vowel formants. *International Journal of Speech, Language and the Law,* 11, pp. 83-102.

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification. *International Journal of Speech, Language and the Law,* 14(2), pp. 224-243.

Campbell, J. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE,* 85 (9), pp. 1437-1462.

Campbell, J., Shen, W., Campbell, W., Schwartz, R., Bonastre, J. F., and Matrouf, D. (2009). Forensic speaker recognition: A need for caution. *IEEE Signal Processing Magazine*, March 2009, pp. 95-103.

Cao, H. and Wang, Y. (2011). A forensic aspect of articulation rate variation in Chinese. *Proceedings of the 17th International Congress of Phonetic Sciences*, August 17-21, 2011, Hong Kong, China, pp. 396-399.

Chambers, J. K. (2005). *Sociolinguistic theory* (2$^{nd}$ ed). Oxford, UK: Blackwell.

Champod, C. and Evett, I. W. (2000). Commentary on Broeders 1999. *Forensic Linguistics,* 7(2), pp. 238-243.

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech communication,* 31, pp. 193-203.

Clark, J., & Yallop, C. (2001). *An Introduction to Phonetics and Phonology* (2nd ed.). Oxford: Blackwell Publishers Ltd.

Clermont, F., French, P., Harrison, P., and Simpson, S. (2008). Population data for English spoken in England: A modest first step*. Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*. Plymouth, United Kingdom.

Clermont, F., Harrison, P., and French, P. (2007). Formant-pattern estimation guided by cepstral compatibility. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*. Plymouth, United Kingdom.

Coe, P. (2012). *The effect of sample contemporaneity on the outcome of likelihood ratios for forensic voice comparison.* Unpublished; University of York. MSc.

Crystal, D. (1987). *The Cambridge Encyclopaedia of Language*. Cambridge: Cambridge University Press.

Damasté, P. H. (1964). Virilization of the voice due to anabolic steroids. *Folia Phoniatrica,* 16, pp. 10-18.

Damasté, P. H. (1967). Voice change in adult women caused by virilizing agents. *Journal of Speech and Hearing Disorders,* 32, pp. 126-132.

Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics,* 25, pp. 287-312.

Darby, J. K. and Hollien, H. (1977). Vocal and speech patterns of depressive patients. *Folia Phoniatrica,* 29, pp. 279-291.

DeGroot, M. and Fienberg S. (1983). The Comparison and Evaluation of Forecasters*. Journal of the Royal Statistical Society*. Series D (The Statistician), Vol. 32 (1), pp. 12-22.

de Jong, G., McDougall, K., and Nolan, F. (2007) Sound Change and Speaker Identity: An Acoustic Study. In: Christian Müller (ed.), *Speaker Classification II: Selected Papers*. Berlin: Springer, pp. 130-141.

Dieroff, H. G. and Siegert, C. (1966). Tonhöhenverschiebung unter Lärmbelastung. *Folia Phoniatrica,* 24, pp. 247-255.

Doherty, R. and Lee, A. (2009). Speech rates of Irish English-speaking adults. *The Royal College of Speech and Language Therapists Scientific Conference* (RCSLT), London, UK.

Drygajlo, A. (2007). Forensic automatic speaker recognition. *IEEE Signal Processing Magazine,* 24, pp. 132-135

Eckert, P. (2000). *Linguistic Variation as Social Practice*. Oxford, UK: Blackwell.

Enzinger, E. (2010a). Characterizing formant tracks in Viennese diphthongs for forensic speaker comparison. *Proceedings of the AES 39th International Conference – Audio Forensics*, Hillerød, Denmark, pp. 47–52.

Enzinger, E. (2010b). Measuring the Effects of the Adaptive Multi-Rate (AMR) codecs on formant tracker performance. *Paper presented at the conference of the Second Pan-American/Iberian Meeting of the Acoustics for the Acoustical Society of America*, Cancún, México.

Enzinger, E. and Morrison, G. S. (2012). The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, pp. 137-140.

Evett, I. W. (1990). The theory of interpreting scientific transfer evidence. *Forensic Science Progress,* 4, pp. 141–179.

Evett, I. W. (1995). Avoiding the transposed conditional. *Science and Justice,* 35 (2), pp. 127-131.

Evett, I. W., Gill, P. D., Jackson, G., Whitaker, J., and Champod, C. (2002). Interpreting small quantities of DNA: The hierarchy of propositions and the use of Bayesian networks*. Journal of Forensic Sciences*, 47(3), pp. 520-530.

Evett, I. W., Jackson, G. Lambert, J. A., and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice,* 40, pp. 233–239.

Evett, I. and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, Massachusetts: Sinauer Associates.

Foulkes, P. and French, P. (2012). Forensic phonetic speaker comparison. In Solan, L. and Tiersma, P. (eds.) *Oxford Handbook of Language and Law*. Oxford: Oxford University Press, pp. 557-572.

French, P. (1990). Acoustic Phonetics. In Baldwin, J. and French, P. (eds.) *Forensic Phonetics*. London: Pinter Publishers. pp. 42-64.

French, J. P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics,* 5(1), pp. 58-68.

French, P., Foulkes, P., Harrison, P., and Stevens, L. (2012). Vocal tract output measures: relative efficacy, interrelationships and limitations. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics Conference*, Santander, Spain.

French, J. P. and Harrison, P. (2007). Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law,* 14(1), pp. 137-144.

French, P. and Harrison, P. (2010). The work of speech and audio analysts. *AES Conference no. 128*. London, United Kingdom.

French, J. P., Nolan, F., Foulkes, P., Harrison, P., and McDougall, K. (2010). The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law*, 17(1), pp. 143-152.

French, P. and Stevens, L. (2013). Forensic speech science. In Jones, M. & R. Knight (eds.), *The Bloomsbury Companion to Phonetics*. London: Bloomsbury, pp. 183-197.

Fritzell, B., Sundberg, J., and Strange-Ebbesen, A. (1982). Pitch change after stripping oedematous vocal folds. *Folia Phoniatrica,* 34, pp. 29-32.

Garrett, K. L. and Healey, E. Ch. (1987). An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day. *Journal of the Acoustical Society of America,* 82, pp. 58-62.

Gilbert, H. R. and Weismer, G. G. (1974). The effects of smoking on the speaking fundamental frequency of adult women. *Journal of Psycholinguistic Research,* 3, pp. 225-231.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics,* 15(2), pp. 87–105.

Giles, H. and Ogay, T. (2007). Communication accommodation theory. In Whaley, B. B. & W. Samter (eds.), *Explaining communication: Contemporary theories and exemplars.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 293-310.

Gimson, A. C. (1970). *An Introduction to the Pronunciation of English*. Second Edition. London: Edward Arnold.

Goel, V. (2000). Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 14 (2), pp. 115-135.

Gold, E. (2009). *The Effects of Video and Voice Recorders in Cellular Phones on Vowel Formants and Fundamental Frequency*. Unpublished; University of York. MSc.

Gold, E. and Hughes, V. (2012). Defining interdependencies between speech parameters. *BBfor2 Short Summer School in Forensic Evidence Evaluation and Validation*. Madrid, Spain.

Gold, E. and Hughes, V. (2013). *Identifying correlations between speech parameters for forensic speaker comparisons*. Research grant provided by the International Association of Forensic Phonetics and Acoustics. [http://www.iafpa.net/GoldHughes_CorrelationsBetweenSpeech Parameters.html]

Goldman-Eisler, F. (1956). The Determinants of the Rate of Speech Output and their Mutual Relations. *Journal of Psychosomatic Research, Vol. 1,* pp 137-143.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press: London.

Gonzalez-Rodriguez, J., Rose, P., Ramos D., Toledano, D. T., and Ortega-Garcia, J. (2007). Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, *IEEE Transactions of Audio, Speech and Language Processing*, 15, pp. 2104-2115.

Graddol, D. (1986). Discourse specific pitch behavior. In Johns-Lewis, C. (ed.). *Intonation in Discourse*. London: Croom Helm, pp. 221-237.

Greenberg, J. H. (1950). Studies in African linguistic classification VI: the click languages. *Southwestern Journal of Anthropology,* 6 (3), pp. 223-237.

Greenberg, C. S., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G., and Godfrey, J. (2010). Human Assisted Speaker Recognition In SRE10, *Proc. Odyssey 2010*, Brno, Czech Republic, June-July 2010.

Grigoras, C. (2001). *Digital Voice Processing System*. Unpublished; University of Bucharest. PhD.

Grigoras, C. (2003). Voice analysis on noisy recordings. *Paper presented at the Cambridge Forensic Phonetics Workshop*, August, Cambridge, UK.

Guillemin, B. and Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language and the Law*, 15(2), pp. 193-218.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Hand, D. J., and Yu, K. (2001). Idiot's Bayes - not so stupid after all? *International Statistical Review,* 69(3), pp. 385-399.

Hardcastle, W. J. (1975). Some aspects of speech production under controlled conditions of oral anaesthesia and auditory masking. *Journal of Phonetics,* 3, pp. 197-214.

Harrison, P. (2004). *Variability of Formant Measurements*. Unpublished; University of York. MA.

Hecker, M. L., Stevens, K. N., von Bismarck, G., and Williams, C. E. (1968). Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America,* 44, pp. 993-1001.

Henze, R. (1953). Experimentelle Untersuchungen zur Phänomenologie der Sprechgeschwindigkeit. *Zeitschrift für Experimentelle und Angewandte Psychologie,* 1, pp. 214-243.

Herbert, R.K. (1990) The sociohistory of clicks in Southern Bantu. *Anthropological Linguistics,* 32 (3), pp. 295-315.

Hodgson, D. (2002) A lawyer looks at Bayes' theorem. *The Australian Law Journal,* 76, pp. 109 – 118.

Hollien, H. (1980). Vocal indicators of psychological stress. In Wright, F., Bahn, C., and Rieben, R.W. (eds.) *Forensic Psychology and Psychiatry*. New York: John Wiley & Sons Ltd., pp. 47-72.

Hollien, H. and Majewski, W. (2009). Unintended Consequences: Due to Lack of Standards for Speaker Identification and Other Forensic Procedures. *Paper presented at the Sixteenth International Congress on Sound and Vibration*, Krakow, Poland, pp. 1-6

Hollien, H. and Michel, J. F. (1968). Vocal fry as a phonational register. *Journal of Speech Language, and Hearing Research,* 11, pp. 600-604.

Horii, Y. (1975). Some statistical characteristics of voice fundamental frequency and chronological age in males. *Journal of Speech, Language, and Hearing Research,* 18, pp. 192-201.

House of Commons Northern Ireland Affairs Committee (2009) *Cross-border Co-operation between the Governments of the United Kingdom and the Republic of Ireland: Second Report of Session 2008–2009*. London: The Stationery Office.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., and Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. *In 16th Proceedings of the International Congress of Phonetic Sciences, Saarbrücken*, pp. 1809-1812.

Hudson, A. I. and Holbrook, A. (1981). A study of the reading fundamental vocal frequency of young black adults. *Journal of Speech, Language, and Hearing Research,* 24, pp. 197-201.

Hughes, V. (in progress*). The Effects of Variability on the Outcome of Numerical Likelihood Ratios*. Unpublished; University of York. PhD.

Hughes, V. (2011). *The Effect of Variability on the Outcome of Likelihood Ratios*. Unpublished; University of York. MSc.

Hughes, V. (2013). Establishing typicality: a closer look at individual formants. In *Proceedings of Meetings on Acoustics*, POMA Volume 19, pp. 060042-060050.

Hughes, V. and Foulkes, P. (2012). Effects of variation on the computation of numerical likelihood ratios for forensic voice comparison. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*. Universidad Internacional Menéndez Pelayo, Santander.

Hughes, V., McDougall, K. and Foulkes, P. (2009). Diphthong dynamics in unscripted speech. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics.* Cambridge, UK. 2-5 August 2009.

Ishihara, S. and Kinoshita, Y. (2008). How many do we need? Exploration of the Population Size Effect on the performance of forensic speaker classification. *9th Annual Conference of the International Speech Communication Association (Interspeech)*. Brisbane, Australia, pp. 1941-1944.

Jacewicz, E., Fox, R. A., O'Neil, C., and Salmons, J. (2009) Articulation rate across dialect, age and gender. *Language Variation and Change,* 21, pp. 233-256.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *Forensic Linguistics*, 4(1), pp. 84-103.

Jessen, M. (2007a). Speaker Classification in Forensic Phonetics and Acoustics. In Müller, C. (ed.), *Speaker classification I: Fundamentals, features, and methods*, Berlin, Germany: Springer, pp. 180-204.

Jessen, M. (2007b). Forensic reference data on articulation rate in German. *Science and Justice,* 47, pp. 50–67.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2(4), pp. 671-711.

Jessen, M. and Becker, T. (2010). Long-term formant distribution as a forensic-phonetic feature. *Paper presented at the Conference of the Acoustical Society of America*. Cancun, Mexico.

Jessen, M. and Roux, J.C. (2002). Voice quality differences associated with stops and clicks in Xhosa. *Journal of Phonetics,* 30, pp. 1–52.

Jessen, M., Enzinger, E., and Jessen, M. (2013). Experiments on Long-Term Formant Analysis with Gaussian Mixture Modeling using VOCALISE. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*, Tampa, Florida, USA.

Jessen, M., Köster, O., and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language, and the Law,* 12(2), pp. 174-213.

Johnson, K. (2003). *Acoustic and auditory phonetics*. (2nd edn). Malden, MA: Blackwell.

Junqua, J. (1996) The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication,* 20, pp. 13-22.

Kavanagh, C. (2010). Speaker discrimination using English nasal durations and formant dynamics. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*. Trier, Germany.

Kavanagh, C. (2011). Intra- and inter-speaker variability in duration and spectral properties of English /s/. *Paper presented at the Conference of the Acoustical Society of America*, San Diego, USA.

Kavanagh, C. (2013). *New Consonantal Acoustic Parameters for Forensic Speaker Comparison*. Unpublished; University of York. PhD.

Kaye, D. H. (1993). DNA evidence: Probability, population genetics, and the courts. *Harvard Journal of Law and Technology*, 7(1), pp. 101-172.

Keating, P. (1988). *A Survey of Phonological Features*. Indiana University Linguistics Club, Bloomington.

Keilmann, A. and Hülse, M. (1992). Dysphonie nach Strumektomie bei ungestörter respiratorischer Beweglichkeit der Stimmlippen. *Folia Phoniatrica,* 44, pp. 261-268.

Kinoshita, Y. (2001). *Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio-Based Approach Using Formants*. Unpublished; Australian National University. PhD.

Kinoshita, Y. (2002). Use of likelihood ratio and Bayesian approach in forensic speaker identification. *Proceedings of the 9th Australian International conference on Speech Science and Technology*. Melbourne, Australia, pp. 297-302.

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term F0. *International Journal of Speech, Language and the Law,* 12, pp. 235-254.

Kinoshita, Y. and Ishihara, S. (2012). The effect of sample size on the performance of likelihood ratio-based forensic voice comparison. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*. Macquarie University, Australia.

Kinoshita, Y., Ishihara, S., and Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language and the Law,* 16, pp. 91-111.

Kinoshita, Y. and Osanai, T. (2006). Within-speaker variation in diphthongal dynamics: what can we compare? *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, University of Auckland, New Zealand, pp. 112-117.

Kirchhübel, C. (2010). The effects of Lombard speech on vowel formant measurements. *Paper presented at the Conference of the Acoustical Society of America*, Cancun, Mexico.

Kirchhübel, C. and Howard, D. (2011). Investigating the acoustic characteristics of deceptive speech. *Proceedings of the 17th International Congress of Phonetic Sciences*, August 17-21, 2011, Hong Kong, China, pp. 1094-1097.

Klingholz, F., Penning, R., and Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from speech signal. *Journal of the Acoustical Society of America,* 84, pp. 929-935.

Kononenko, I. (1990). Comparison of inductive and naïve Bayesian capitalised learning approaches to automatic knowledge acquisition. In B. Wielinga et al. (Eds.) *Current trends in knowledge acquisition*. IOS Press: Amsterdam, Netherlands, pp. 190-197.

Konrat, C. and Jessen, M. (2013). Fundamental Frequency Analysis: A Collaborative Exercise. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*, Tampa, Florida, USA.

Koreman, J. (2006). Perceived speech rate: the effects of articulation rate and speaking rate in spontaneous speech. *Journal of the Acoustical Society of America,* 119, pp. 582-596.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Künzel, H. J. (1987). *Sprechererkennung: Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag.

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica*, 46, p. 117.

Künzel, H. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics,* 4, pp. 48-83.

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, 7, pp. 149-179.

Künzel, H. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law,* 8, pp. 80-99.

Künzel, H., Braun, A., and Eysholdt, U. (1992). *Einfluβ von Alkohol auf Stimme und Sprache*. Heidelberg: Kriminalistik-Verlag.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladefoged, P. (2006). *A Course in Phonetics*, 5th edn. Boston: Wadsworth Cengage Learning.

Laplace, P. S. (1781). Mémoire sur les Probabilitiés. *OC* (9), pp. 383-485.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.

Law Commission of England and Wales (2011). Expert Evidence in Criminal Proceedings in England and Wales. No. 325

Liénard, J. and Di Benedetto, M. (199). Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America,* 106(1), pp. 411-422.

Lindh, J. (2006) 'Preliminary Descriptive F0-statistics for Young Male Speakers' *Lund University Working Papers,* 52, pp. 89-92.

Lindh, J. (2007). Fundamental frequency and the alternative baseline in forensic speaker identification. *Paper presented at the Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, Plymouth, UK.

Lindh, J., & Morrison, G.S. (2011). Humans versus machine: Forensic voice comparison on a small database of Swedish voice recordings. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, pp. 1254–1257.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika,* 64, pp. 207-213.

Loakes, D. (2006). Variation in long-term fundamental frequency: measurements from vocalic segments in twins' speech. *Proceedings of the 11th Speech Science and Technology Conference*, Auckland, New Zealand, pp. 205-210.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des maladies de l'oreille et du larynx,* 37, pp. 101-119.

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists.* Chichester: John Wiley & Sons, Ltd.

Mead, K. O. (1974). *Identification of speakers from fundamental-frequency contours in conversational speech*. Joint Speech Research Unit, Report No. 1002. Ruislip, Middlesex.

Miller, J. L., Grosjean, F., and Lomanti, C. (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. *Phonetica,* 41, pp. 215-225.

McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law,* 11 (1), pp. 103-130.

McDougall, K. (2006). The role of formant dynamics in determining speaker identity. Ph.D. abstract. *International Journal of Speech, Language and the Law*, 13(1), pp. 144-145.

McDougall, K. and Nolan, F. (2007). Discrimination of speakers using formant dynamics of /u:/ in British English. *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany, pp. 1445-1448.

Milroy, L. and Gordon, M. (2003). *Sociolinguistics: Method and Interpretation.* Oxford: Blackwell.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician,* 101, pp. 7-24.

Morrison, G. S. (2007). MatLab implementation of Aitken and Lucy's (2004) forensic likelihood ratio software using multivariate-kernel-density estimation. Downloaded: December 2011.

Morrison, G. S. (2008) Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language, and the Law,* 15(2), pp. 249-266.

Morrison, G. S. (2009a). Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America,* 125, pp. 2387-2397.

Morrison, G. S. (2009b). The place of forensic voice comparison in the ongoing paradigm shift. Written version of an invited presentation given at the 2nd International Conference on Evidence Law and Forensic Science. 25-26 July, Beijing, China, pp. 1-16.

Morrison, G. S. (2009c). Forensic voice comparison and the paradigm shift. *Science and Justice*, 49, pp. 298-308.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication,* 53, pp. 242-256.

Morrison, G. S. (2012) Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45 (2), pp.173-197.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences,* 45, pp. 173-197.

Morrison, G. S. and Kinoshita, Y. (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. *Proceedings of Interspeech 2008 International Speech Communication Association*, pp. 1501–1504.

Morrison, G. S., Ochoa, F., and Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore, International Speech Communication Association*, pp. 62-77.

Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences,* 44, pp. 155–167.

Morrison, G.S., Thiruvaran, T., and Epps, J. (2010). An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*. Melbourne, Australia, pp. 74-77.

Murphy, C. H. and Doyle, P. C. (1987). The effects of cigarette smoking on voice-fundamental frequency. *Otolaryngology-Head and Neck Surgery,* 97, pp. 376-380.

Narang, V., Misra, D. and Yadav, R. (2012). F1 and F2 correlation with F0: A study of vowels of Hindi, Punjabi, Korean, and Thai. *International Journal of Asian Language Processing,* 22(2), pp. 63-73.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington D.C.: The National Academic Press.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nolan, F. and Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law,* 12(2), pp. 143-173.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law,* 16(1), 31-57.

Novak, A., Dlouha, O., Capkova, B., and Vohradnik, M. (1991). Voice fatigue after theater performance in actors. *Folia Phoniatrica,* 43, pp. 74-78.

Oats, J. M. and Dacakis, G. (1983). Speech pathology considerations in the management of transsexualism – a review. *British Journal of Disorders of Communication,* 18, pp. 139-151.

Ogden, R. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association,* 43(3), pp. 299-320.

Osterburg, J. W. (1969). The evaluation of physical evidence in criminalistics: Subjective or objective process?, *Journal of Criminal Law and Criminology,* 60, pp. 97-101.

Papp, V. (2008). *The Effects of Heroin on Speech*. Unpublished; University of York. MSc.

Peterson, G. E. (1959). The acoustics of speech – part II: acoustical properties of speech waves. In Travis, L. E. (ed.) *Handbook of Speech Pathology*. London: Peter Owen, pp. 137-173.

Pike, K. (1943). *Phonetics: A critical analysis of phonetic theory and a technic for the practical description of sounds*. Ann Arbor, MI: University of Michigan Press.

Pisoni, D. B. and Martin, C. S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analysis. *Alcoholism: Clinical and Experimental Research,* 13, pp. 577-587.

Porwal, U., Shi, Z., and Setlur, S. (2013). Machine learning in handwritten Arabic text recognition. In Govindaraju, V. and C. R. Rao (eds.) *Handbook of Statistics 31 - Machine Learning: Theory and Applications*. Amsterdam: North Holland, pp. 443-470.

Ramos-Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. Unpublished PhD Dissertation, Universidad Autónoma de Madrid.

Ramos-Castro, D., Gonzalez-Rodriguez, J. and Ortega-Garcia, J. (2006). Likelihood Ratio Calibration in a Transparent and Testable Forensic Speaker Recognition Framework. *In the Proceedings of the IEEE Speaker and Language Recognition Workshop*, pp. 1-8.

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. *Acoustics, Speech, and Signal Processing (ICASSP), Proceedings of the IEEE International Conference* 4, pp. 4072-4075.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10 (1-3), pp. 19-41.

Rhodes, R. (2013). *Assessing the Strength of Non-Contemporaneous Forensic Speech Evidence*. Unpublished; University of York. PhD.

Robb, M., Maclagan, M. A., and Chen, Y. (2004) Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics,* 18(1), pp. 1-15.

Robertson, B. and Vignaux, G. A. (1995). *Interpreting evidence: evaluating forensic science in the courtroom*. Chichester: John Wiley & Sons, Ltd.

Rose, P. (1999). Long- and short-term within-speaker differences in the formants of Australian hello. *Journal of the International Phonetic Association,* 29(1), pp. 1-31.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor-Francis Ltd.

Rose, P. (2003). The technical comparison of forensic voice samples. In I.S. Freckleton and H. Selby (eds.) *Expert Evidence*. North Ryde: Lawbook Co, Ch. 99.

Rose, P. (2006a). The intrinsic forensic discriminatory power of diphthongs. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. University of Auckland, New Zealand, pp. 64-69.

Rose, P. (2006b). Technical forensic speaker recognition: evaluation, types and testing of evidence. *Computer Speech and Language,* 20, pp. 159–91.

Rose, P. (2006c). Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination. *Proceedings of the Speaker and Language Recognition Workshop*, pp. 1- 8.

Rose, P. (2007a). Forensic speaker discrimination with Australian English vowel acoustics. *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany, pp. 1817-1820.

Rose, P. (2007b). Going and getting it – Forensic Speaker Recognition from the perspective of a traditional practitioner/researcher. Paper presented at the *Australian Research Council Network in Human Communicative Science Workshop: FSI not CSI – Perspectives in State-of-the-Art Forensic Speaker Recognition*, Sydney.

Rose, P. (2010a). Bernard's 18 – vowel inventory size and strength of forensic voice comparison evidence. *Proceedings of the 13th Australian International Conference on Speech and Technology*. Melbourne, Australia, pp. 30-33.

Rose, P. (2010b). The effect of correlation on strength of evidence estimates in forensic voice comparison: uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics. *International Journal of Biometrics,* 2, pp. 316-329.

Rose, P. (2011). Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra. In W. S. Lee, E. Zee (Eds.) *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 17-21 August, pp. 1718-1721.

Rose, P. (2012). Yes, Not Too Bad — Likelihood Ratio-Based Forensic Voice Comparison in a $150 Million Telephone Fraud. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology.* Macquarie University, Australia, pp. 161-164.

Rose, P. (2013a). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language and the Law,* 20(1), pp. 77-116.

Rose, P. (2013b). Where the science ends and the law begins: Likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *International Journal of Speech, Language and the Law,* 20(2), pp. 277-324.

Rose, P., Kinoshita, Y., and Alderman, T. (2006). Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. University of Auckland, New Zealand, pp. 329-334.

Rose, P., Lucy, D., and Osanai, T. (2004). Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects model: a "non-idiot's Bayes" approach. *Proceedings of the 10th Australasian Conference on Speech Science and Technology*. Macquarie University, Australia, pp. 492-497.

Rose, P. and Morrison, G. S. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law,* 16 (1), pp. 139-163.

Rose, P., Osanai, T., and Kinoshita, Y. (2003). Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Forensic Linguistics*, 10, pp. 179-202.

Rose, P. and Winter, E. (2010). Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio approaches. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*. Melbourne, Australia, pp. 42-45.

Saks, M. J. and Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science,* 309, pp. 892-895.

Saxman, J. H. and Burk, K. W. (1968). Speaking fundamental frequency characteristics of adult female schizophrenics. *Journal of Speech, Language, and Hearing Research,* 11, pp. 194-203.

Scherer, K. R. (1977). The effect of stress on fundamental frequency of the voice. *Journal of the Acoustical Society of America,* 6, pp. 25-26.

Scherer, K. R., Helfrich, H., Standke, R., and Wallbott, H. (1976). *Psychoakustische und kinesische Verhaltensanalyse*. Research report, University of Giessen.

Schultz-Coulon, H. J. (1975). Bestimmung und Beurteilung der individuellen mittleren Sprechstimmlage. *Folia Phoniatrica,* 27, pp. 375-386.

Schultz-Coulon, H. J. and Fues, C. P. (1976). Der Lombard-Reflex als Stimmfunktionsprüfung. *HNO* 24, pp. 200-204.

Shepard, C. A., Giles, H. and LePoire, B. (2001). Communication Accommodation Theory. In, Robinson, W. Peter & Howard Giles (eds.)*, The New Handbook of Language and Social Psychology.* Chichester: Wiley, pp. 34-51.

Simpson, S. (2008). *Testing the Speaker Discrimination Ability of Formant Measurements in Forensic Speaker Comparison Cases*. Unpublished; University of York. MSc.

Sobell, L. C., Sobell, M. B., and Coleman, R. F. (1982). Alcohol-induced dysfluency in nonalcoholics. *Folia Phoniatrica,* 34, pp. 316-323.

Sorensen, D. and Horii, Y. (1982). Cigarette smoking and voice fundamental frequency. *Journal of Communication Disorders,* 15, pp. 135-144.

Steffan-Batog, M., Jassem, W., and Gruszka-Koscielak, H. (1970). Statistical distributions of short term F0 values as a personal voice characteristic. In Jassem, W. (ed.) *Speech Analysis and Synthesis*. Warsaw: Polish Academy of Sciences, Volume 2: 197-208.

Stevens, K. N. (2001). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, L. and French, P. (2012). Voice quality in standard southern British English: distribution of features, inter-speaker variability, and the effect of telephone transmission. *Paper presented at the Conference of the International Association of Forensic Phonetics and Acoustics*. Santander, Spain.

Stoney, D. A. (1991). What made us ever think we could individualize using statistics? *Journal of the Forensic Science Society,* 31(2), pp. 197

SurveyGizmo (2006). *International Survey on Forensic Speaker Comparison Practices*. http://www.surveygizmo.com Accessed 21 July, 2010.

Syrdal, A. and Steele, S. (1985). Vowel F1 as a function of speaker fundamental frequency. *Journal of the Acoustical Society of America,* 78, S56.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: the Prosecutor's fallacy and the Defence attorney's fallacy. *Law and Human Behaviour,* 11 (3), pp. 167-187.

Traunmüller, H. and Eriksson, A. (1995). The Frequency Range of the Voice Fundamental in the Speech of Male and Female Adults (unpublished manuscript) Visited August 9, 2013 from: http://www2.ling.su.se/staff/hartmut/f0_m%26f.pdf.

Trouvain, J. (2004). Tempo variation in speech production. Implications for speech synthesis. Doctoral dissertation published as Report Nr. 8 of *Reports in Phonetics*, University of the Saarland.

Trudgill, P. (1981). Linguistic accommodation: Sociolinguistic observations on a sociopsychological theory. In Hendrick, R., Mase, C. and Miller, M. F. (eds.), *Papers from the Parasession on Language and Behavior*. Chicago, IL.: Chicago Linguistic Society, pp. 193-247.

U.S. National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington D.C.: The National Academies Press.

van Leeuwen, D. A. and Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In Müller, C. (ed.) *Speaker Classification I*, LNAI 4343. Berlin: Springer-Verlag, pp. 330-353.

Vermeulen, J. (2009). *Beware of the Distance: Evaluation of Spectral Measurements of Synthetic Vowels Recorded at Different Distances.* Unpublished; University of York. MSc.

Wardhaugh, R. (2006). *An Introduction to Sociolinguistics* (5th ed.). Oxford, UK: Blackwell.

Williams, C. E. and Stevens, K. N. (1972). Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America,* 52, pp. 1238-1250.

Wright, M. (2005). *Studies of the Phonetic-Interaction Interface: Clicks and Interactional Structures in English Conversation*. Unpublished; University of York. PhD.

Wright, M. (2007). Clicks as markers of new sequences in English conversation. *Proceedings of the 16th International Congress of the Phonetic Sciences,* Saarbrücken *(*ICPhS *XVI)*, pp. 1069-1072.

Wright, M. (2011a). On clicks in English talk-in-interaction. *Journal of the International Phonetic Association*, 41(2), pp. 207-229.

Wright, M. (2011b). The phonetics–interaction interface in the initiation of closings in everyday English telephone calls. *Journal of Pragmatics*, 43(4), pp. 1080-1099.

Xue S. A. and Hao J. G. (2006). Normative Standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *Journal of Voice,* 20, pp. 391-400.

Zetterholm, E. (2006). Same speaker – different voices: A study of one impersonator and some of his different imitations. In *Proceedings of the 11th Speech Science and Technology Conference*, Auckland, New Zealand, pp. 70-75.

Zhang, C., Morrison, G. S., Rose, P. (2008). Forensic speaker recognition in Chinese: a multivariate likelihood ratio discrimination on /i/ and /y/. *Proceedings of Interspeech 2008: Incorporating SST 2008 and International Speech Communication Association*, pp. 1937–1940.

## COURT CASES

*Daubert* v *Merrell Dow Pharmaceuticals, Inc.* (509 US 579 [1993]).

*Frye* v *United States* (293 F. 1013 D.C. Cir. [1923])

*Joseph Crossfield and Sons Ltd* v *Techno Chemical Laboratories Ltd* (29 TLR, 378 & 379 [1913])

*Kumho Tire* v *Carmichael* (526 U.S. 137 [1999])

*R* v *Bonython* [1984] 38 SASR 45

*R* v *Doheny and Adams* [1996] EWCA Crim 728

*R* v *Flynn and St John* [2008] EWCA Crim 970

*R* v *Hufnagl* [2008] NSWDC 134

*R* v *Robb* [1991] 93 EWCA Crim 161

*R* v *T* [2011] EWCA Crim 729

*United States* v *Starzecpyzel* (880 F. Supp. 1027 S.D.N.Y. [1995])