Forensic Comparison of Voices, Speech and Speakers

Tools and Methods in Forensic Phonetics

Jonas Lindh

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg

UNIVERSITY OF GOTHENBURG

Doctoral Dissertation in linguistics
University of Gothenburg
2017-06-07

Original cover art: Jens Åkesson (*“Fiction-troublemaker”*)
Design artwork: Joel Åkesson

Printed by Reprocentralen Lorensberg

ISBN 978-91-629-0141-7 (print)
ISBN 978-91-629-0142-4 (digital)

e-publication available at:
http://hdl.handle.net/2077/52188

Distribution:
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Box 200, SE-40530 Gothenburg

Acknowledgments

Sometimes a thank you is not enough and sometimes it is. It does not matter how many names one acknowledges, someone will probably always be forgotten. Possibly because they were rightfully not on the top of the list or simply due to negligence of the author. Not because he is spiteful, but because his hardware space is too small.

The beginning of my work was supervised by professor emeritus Anders Eriksson, which I am grateful for. I would like to thank my supervisor professor Åsa Abelin for her advice and encouragement throughout this long and difficult project. My co-supervisor professor Paul Foulkes cannot be thanked enough for his absolutely invaluable expert advice, encouragement and long discussions, both formal and informal ones. I would also like to thank Geoffrey Morrison for valuable input regarding many things in this work that has puzzled my mind. From my early years of studying phonetics I would like to thank Per Lindblad, the most inspiring teacher I have ever met.

I am extremely grateful for the support from the Graduate School of Language Technology (GSLT) that I received in terms of courses, supervision, ideas, retreats, travel funding, colleagues and friends. No one mentioned by name and no one forgotten. Being associated with the GSLT has not only provided me with funding, great colleagues with a wide range of knowledge and input, but also a huge national and international network of people to ask for guidance and/or cooperation. My PhD student colleagues from the Department of Linguistics, and more recently, Philosophy, Linguistics and Theory of Science, have all supported me with their comments, discussions and encouraging cheering. Thank you. I would like to send my gratitude to all my colleagues at Audiology and Speech and Language Pathology, Sahlgrenska Academy.

Thank you for your invaluable support Joel. Without you as mental coach, friend and squire in good and bad times I would have never been able to get this work done. Thank you Catherine McHale Gunnarsson for providing me very valuable language support. I would like to thank my parents for being supportive most of the time, even when they could not understand at all what I was doing. My four children and wife is of course always in the center of my attention. Unfortunately I have at times forgotten that and neglected the importance of their support and the efforts involved in leaving me working long hours in the office or in front of a screen on the kitchen table. Without them, the author's life would have been very poor and this work would have never been concluded.

Abstract

Ph.D. dissertation at University of Gothenburg, Sweden, 2017

Title: *Forensic Comparison of Voices, Speech and Speakers*
Author: Jonas Lindh
Language: English, with a Swedish summary
Department: Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Box 200, SE-405 30 Göteborg
ISBN: 978-91-629-0142-4 (digital)
ISBN: 978-91-629-0141-7 (print)

This thesis has three main objectives. The first objective (A) includes Study I, which investigates the parameter fundamental frequency (F0) and its robustness in different acoustic contexts by using different measures. The outcome concludes that using the alternative baseline as a measure will diminish the effect of low-quality recordings or varying speaking liveliness. However, both creaky voice and raised vocal effort induce intra-variation problems that are yet to be solved.

The second objective (B) includes study II, III and IV. Study II investigates the differences between the results from an ear witness line-up experiment and the pairwise perceptual judgments of voice similarity performed by a large group of listeners. The study shows that humans seem to be much more focused on similarities of speech style than features connected to voice quality, even when recordings are played backwards. Study III investigates the differences between an automatic voice comparison system and humans' perceptual judgments of voice similarity. The experiments' results show that it is possible to see a correlation between how speakers were judged as more or less different using multidimensional scaling of similarity ranks compared to both the automatic system and the listeners. However, there are also differences due to the fact that human listeners include information about speech style and have difficulties weighting the parameters, i.e. ignoring them when they are contradictory. Study IV successfully investigates a new functional method for how to convert the perceptual similarity judgments made by humans and then compare those to the automatic system results within the likelihood ratio framework. It was discovered that the automatic system outperformed the naive human listeners in this task (using a very small dataset).

The third objective (C) includes study V. Study V investigates several statistical modelling techniques to calculate relevant likelihood ratios using simulations based on existing reference data in an authentic forensic case of a disputed utterance. The study presents several problems with modelling small datasets and develops methods to take into account the lack of data within the likelihood ratio framework.

In summary, the thesis contains a larger historical background to forensic speaker comparison to guide the reader into the current research situation within forensic phonetics. The work further seeks to build a bridge between forensic phonetics and automatic voice recognition. Practical casework implications have been considered throughout the work on the basis of own experience as a forensic caseworker and through collaborative interaction with other parties working in the field, both in research and in forensic practice and law enforcement. Since 2005, the author has been involved in over 400 forensic cases and given testimony in several countries.

Keywords: forensic phonetics, automatic voice recognition, disputed utterance, speech, language technology

Sammanfattning (svenska)

Titel (engelska): *Forensic Comparison of Voices, Speech and Speakers*
Titel (svenska): *Forensiska jämförelser av röster, tal och talare*
Författare: Jonas Lindh
Språk: Engelska (med svensk sammanfattning)
Institution: Institutionen för filosofi, lingvistik och vetenskapsteori, Göteborgs universitet
ISBN: 978-91-629-0142-4 (digital)
ISBN: 978-91-629-0141-7 (print)

Denna sammanläggningsavhandling har tre huvudmål. Det första huvudmålet (A), innefattar studie I, undersöker parametern grundtonsfrekvens (F0) och dess stabilitet i olika akustiska och lingvistiska kontexter med hjälp av olika mått. Resultatet visar att användningen av den så kallade alternativa baslinjen kommer att minska effekten av olika inspelningar med låg kvalitet och varierande livlighetsnivå av talet. Både knarrig röst och varierad röststyrka ger dock variationsproblem som återstår att lösa.

Det andra huvudmålet (B), innefattar studie II, III och IV. Studie II undersöker skillnaderna mellan resultaten från ett öronvittnesexperiment och parvisa perceptuella bedömningar av röstlikhet som utförts av en stor grupp lyssnare. Studien visar att människor verkar vara mycket mer fokuserade på likheterna mellan talstil än parametrar kopplade till röstkvalitet, även när inspelningar spelas baklänges. Studie III undersöker skillnaderna mellan ett automatiskt röstjämförelsesystem och människors perceptuella bedömningar av röstlikhet. Experimentets resultat visar att det är möjligt att se ett samband mellan hur talare bedömdes som mer eller mindre olika med hjälp av multidimensionell skalning. Men det visade sig också finnas skillnader mellan perceptuella bedömningar och det automatiska systemets resultat. Dessa verkar bero på det faktum att lyssnare mer använder information om talstil i sina bedömningar och det textoberoende automatiska systemet enbart röstkvalitetsaspekter. Studie IV undersöker framgångsrikt en ny metod för hur man kan omvandla de perceptuella röstlikhetsbedömningarna gjorda av människor på en ordinalskala till likelihood-kvoter likt resultatet från ett automatiskt system. Detta för att sedan bättre kunna jämföra dem med det automatiska systemets resultat. Det upptäcktes också att det automatiska systemet hade bättre diskriminationsresultat mellan rösterna än de mänskliga lyssnarna i denna uppgift (med en mycket liten testdatabas).

Det tredje huvudmålet (C) innefattar studie V. Studie V undersöker flera statistiska modelleringsmetoder för att beräkna relevanta likelihood-kvoter med hjälp av simuleringar. Simuleringarna är baserade på befintliga referensdata från ett autentiskt rättsfall där en analys av ett forensiskt omtvistat yttrande användes. Studien presenterar flera problem med modellering av små datamängder och utvecklar metoder för att ta hänsyn till bristen på data vid uträkningar av likelihood-kvoter.

Sammanfattningsvis innehåller avhandlingen också en större historisk bakgrund till forensisk röstjämförelse för att guida läsaren i det aktuella forskningsläget inom rättsfonetik. Arbetet syftar vidare till att bygga en bro mellan rättsfonetik och automatisk röstigenkänning. Under arbetet med avhandlingen har rättsfonetisk praktik legat till grund för många tankar och idéer. Detta har kommit naturligt, speciellt med tanke på författarens egna erfarenheter som rättsfonetisk analytiker och genom den samverkan med både forskning och rättsväsendet detta inneburit. Sedan 2005 har författaren fungerat som sakkunnig i mer än 400 rättsfonetiska ärenden och vittnat i flertalet länder.

Nyckelord: rättsfonetik, automatisk röstigenkänning, omtvistat yttrande, tal, röst, talteknologi

1 TABLE OF CONTENTS

1 TABLE OF CONTENTS **1**

2 PUBLICATIONS AND CONTRIBUTORS **3**

2.1 Non-included papers related to the present thesis **4**

3 LIST OF ABBREVIATIONS **7**

4 BACKGROUND AND INTRODUCTION **9**

4.1 General overview **9**

4.1.1 A likely scenario for a forensic speaker-comparison case 10

4.1.2 A likely scenario for a forensic disputed utterance case 12

4.1.3 A likely scenario for an ear-witness line-up 12

4.2 General introduction to the included studies **13**

4.3 Historical Background **16**

4.3.1 The Voiceprint Controversy 18

4.3.2 Forensic phonetics in Sweden 26

4.4 Short Review of Forensic Phonetics **27**

4.4.1 A classical phonetician's approach 27

4.4.2 An engineer's approach 28

4.4.3 A possible new generation 30

4.5 Modern Forensic Speaker Comparison **31**

4.5.1 Short review of acoustic measurements and automatic systems 32

4.5.2 Summary 33

5 RESEARCH FOR THE THESIS **34**

5.1 Objectives **34**

5.2 Research questions **35**

5.2.1 Fundamental frequency from an individual perspective 35

5.2.2 Human perception and Automatic Voice Comparison from a forensic perspective 35

5.2.3 Forensic transcription and disputed utterances 37

5.3 Various general methods and definitions **37**

5.3.1 Robustness of humans and machines 37

5.3.2 The concept of robustness 38

5.3.3 Robustness of forensic phonetic parameters 40

5.4 Tools and databases **47**

5.4.1 The Terminology for automatic systems 48

5.4.2 Parameterization 49

5.4.3 General Performance and Evaluation of Automatic Systems 53

5.4.4 ALIZE - an open-source platform for biometrics 56

5.4.5 Plugin and Database 59

5.4.6 UBM Database 61

6 GENERAL CONTRIBUTIONS OF THE THESIS **65**

6.1 Summary of Findings **65**

6.1.1 Study I 65

6.1.2 Study II 66

6.1.3 Study III 67

6.1.4 Study IV 68

6.1.5 Study V 68

7 GENERAL CONCLUSIONS **71**

7.1 Objective A **71**

7.2 Objective B **72**

7.3 Objective C **73**

8 APPLICATIONS AND FUTURE RESEARCH **74**

8.1 Applications in casework **74**

8.1.1 The use of a robust F0 parameter in casework – Study I 74

8.1.2 Combination of AVC and Perceptual Analysis – Studies II, III, IV 75

8.1.3 Cases of Disputed Utterances Analysis – Study V 75

8.2 Outlook **76**

9 REFERENCES **78**

Appendix - Attached Papers **104**

2 PUBLICATIONS AND CONTRIBUTORS

Some of the included studies have been conducted in collaboration with others. The details of these collaborations are specified below:

Study I. Lindh, J. & Eriksson, A. (2007). Robustness of Long Time Measures of Fundamental Frequency, *In Proceedings of Interspeech 2007, Antwerp, Belgium* 2025–2028. ISBN: 9781605603162[1]

Eriksson contributed parts of the data and parts of the statistical analyses, and supervised the writing of the paper.

Study II. Lindh, J. (2009). Perception of voice similarity and the results of a voice line-up, *The XXII Swedish Phonetics Conference, Department of Linguistics, Stockholm University, 2009*. pp. 186-189. ISBN/ISSN: 978-91-633-4892-1

Study III. Lindh, J. & Eriksson, A. (2010). Voice similarity — a comparison between judgements by human listeners and automatic voice comparison, *Proceedings from FONETIK 2010, Working Papers*. 54 pp. 63-69.

Eriksson contributed to parts of the statistical analyses and the supervision of the writing of the paper.

Study IV. Lindh, J. & Morrison, G. S. (2011). Humans versus machine: Forensic voice comparison on a small database of Swedish voice recordings, *Proceedings of ICPhS2011*. [2]

Morrison contributed to ideas about the mathematics and supervised the writing of the paper.

Study V. Morrison, G. S., Lindh, J. & Curran, J. M. (2014). Likelihood ratio calculation for a disputed-utterance analysis with limited available data. *Speech Communication* 58 pp. 81-90.[3]

Morrison and Curran contributed statistical and mathematical ideas and jointly authored the paper.

[1] Study I is reprinted with kind permission from ISCA.

[2] Study II is reprinted with kind permission of the ICPHs.

[3] Study V is reprinted with kind permission of Speech Communication (Elsevier).

2.1 Non-included papers related to the present thesis

Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J., & Åkesson, J. (2016). Identifying Perceptually Similar Voices with a Speaker Recognition System Using Auto-Phonetic Features. *Interspeech 2016*, 1567-1568.

Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J. & Åkesson, J. (2016). *Automatically identifying perceptually similar voices for voice parades.* Presented at IAFPA25, pp. 25-26.

Lindh, J., Åkesson, J. & Sundqvist, M. (2016). *Comparison of Perceptual and ASR Results on the SweEval2016 Corpus.* Poster presented at IAFPA25, pp. 110-111.

Lindh, J. & Åkesson, J. (2016). *Evaluation of Software 'Error checks' on the SweEval2016 Corpus for Forensic Speaker Comparison.* Presented at IAFPA25. pp. 57-58.

Forsberg, J., Gross, J., Lindh, J. & Åkesson, J. (2015). *A forensic and sociophonetic perspective on a new corpus of young urban Swedish.* Poster presented at 10th UK Language Variation and Change (UKLVC) conference 1-3/9 2015, York, UK.

Forsberg, J., Gross, J., Lindh, J. & Åkesson, J. (2015). *Speaker comparison evaluation using a new corpus of urban speech.* Poster presented at the 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics, 8-10/7 2015, Leiden. pp. 46-47.

Lindh, J. (2015). *Forensic speaker comparison evaluations.* Presented at Roundtable in Forensic Linguistics 2015, September 4th- 6th, Mainz, Germany.

Lindh, J. (2015). *Forensic speaker comparison using machine and mind.* Presented at 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics, 8 - 10 July 2015, Leiden.

Lindh, J. & Åkesson, J. (2014). *Effect of the Double-Filtering effect on Automatic Voice Comparison.* Poster presented at IAFPA 2014. International Association for Forensic Phonetics and Acoustics Annual Conference 31 August - 3 September 2014.

Lindh, J. & Åkesson, J. (2013). *A pilot study on the effect of different phonetic acoustic input to a GMM - UBM system for voice comparison.* Poster presented at the 22nd conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). July 21st-24th, 2013, Tampa, Florida, USA.

Åkesson, J. & Lindh, J. (2013). *Describing a database collection procedure for studying 'double filtering' effects.* 22nd conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). July 21st-24th, 2013, Tampa, Florida, USA.

Lindh, J., Ochoa, F. & Morrison, G. S. (2012). *Calculating the reliability of a likelihood ratio from a disputed utterance.* Presented at the 21st conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). August 5th-8th, 2012, Santander, Spain.

Morrison, G. S., Ochoa, F. & Lindh, J. (2012). *Calculating the reliability of likelihood ratios: Addressing modelling problems related to small n and tails.* Presented at the UNSW Forensic Speech Science Conference. 3 December 2012. Sydney, Australia.

Lindh, J., Eriksson, A. & Nelhans, G. (2010). *Methodological Issues in the Presentation and Evaluation of Speech Evidence in Sweden,* Presented at the 19th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Trier, Germany.

Lindh, J. (2010). *Preliminary Formant Data of the Swedia Dialect Database in a Forensic Phonetic Perspective.* Poster presented at the 19th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Trier, Germany.

Lindh, J. (2009). *A first step towards a text-independent speaker verification Praat plug-in using Mistral/ALIZE tools.* In Proceedings of the XXIIth Swedish Phonetics Conference, Department of Linguistics, Stockholm University, 2009, pp. 194-197. ISBN/ISSN: 978-91-633-4892-1

Lindh, J. (2009). *Pick a Voice among Wolves, Goats and Lambs.* Presented at the 18th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Cambridge, UK.

Lindh, J., & Eriksson, A. (2009). The SweDat Project and Swedia Database for Phonetic and Acoustic Research. In *Proceedings of the 2009 Fifth IEEE International Conference on e-Science* (pp. 45–49). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/e-Science.2009.15

Lindh, J. (2008). *Robustness of Forced Alignment in a Forensic Context.* Presented at the 17th Conference of the International Association for Forensic Phonetics and Acoustics, Lausanne, Switzerland.

Lindh, J. (2006). Preliminary Descriptive F0-statistics for Young Male Speakers. In S. Schötz & G. Ambrazaitis (Eds.), *Working Papers 52: Papers from Fonetik 2006*. Lund, Sweden: Department of Linguistics, Lund University, pp. 89-92.

Lindh, J. (2006). *Preliminary F0 Statistics and Forensic Phonetics,* Presented at the 15th Conference of the International Association for Forensic Phonetics and Acoustics, Gothenburg, Sweden. (Eds. Jonas Lindh and Anders Eriksson).

Lindh, J. (2004). *Handling the "Voiceprint" Issue.* In proceedings of the XVIIth Swedish Phonetics Conference, Stockholm, Sweden, pp. 72-75.

3 LIST OF ABBREVIATIONS

AR	Articulation Rate
ASC	Automatic Speaker Comparison
ASR	Automatic Speech Recognition
AVC	Automatic Voice Comparison
DET	Detection Error Trade-off
EER	Equal Error Rate
EM	Expectation Maximization
ENFSI	European Network of Forensic Science Institutes
FSC	Forensic Speaker Comparison
FT	Fourier Transforms
FVC	Forensic Voice Comparison
GMM	Gaussian Mixture Model
GSLT	Graduate School of Language Technology
HASR	Human Assisted Speaker Recognition
IAFP	International Association of Forensic Phonetics
IAFPA	International Association for Forensic Phonetics and Acoustics
IAI	International Association for Identification
IRCGN	Institut de recherche criminelle de la gendarmerie nationale
LPC	Linear Predictive Coding
LR	Likelihood Ratio
LT	Language Technology
LTAS	Long-Term Average Spectrum
LTF	Long-Term Formant
MAP	Maximum a Posteriori
MFC	Mel-Frequency Cepstrum
MFCC	Mel-Frequency Cepstral Coefficients
MIT	Massachusetts Institute of Technology
NFC	National Forensic Centre
NIST	National Institute of Standards and Technology
PLP	Perceptual Linear Predictive

PVI	Pairwise Variability Index
ROC	Receiver Operating Characteristic
SPAAT	Super Phonetic Annotation and Analysis Tool
SPro	Signal Processing Toolkit[4]
SR	Speaking Rate
UBM	Universal Background Model
UKLVC	UK Language Variation and Change
USSS	United States Secret Service
VCS	Voice Comparison Standards
VOT	Voice Onset Time

[4] http://www.irisa.fr/metiss/guig/spro/

4 BACKGROUND AND INTRODUCTION

4.1 General overview

Forensic analysis has for a long time been confused with identification. Forensic analysis is not identification, but comparison between samples under competing hypotheses. Forensic speaker or voice comparison (FSC/FVC) has developed a lot over the last decade with the development of automatic systems and methodologies. Although the terms speaker and voice are often used interchangeably, especially in forensic speech science, in this thesis they are defined distinctly as follows. Voice refers to the holistic acoustic product of the biologically constrained vocal tract and/or vocal folds. What makes a speaker different from another is, however, much more than just the difference in voice: Speaker refers to the linguistic, social, pragmatic, behavioural and idiosyncratic properties that are conveyed through the voice. For example, two voices might be very similar judged perceptually or calculated through an automatic system, but might in fact contain two radically different accents or even languages. The approach taken in this thesis is that it is essential for forensic analysis to investigate potentially many aspects of the speaker, not simply the voice in the holistic acoustic sense. Therefore the term FSC is preferred to FVC. What is referred to as speech in the thesis is the combination of voice and the linguistic, and even non-linguistic, elements encoded in the acoustic signal. The linguistic elements include sound units of the language (allophones), higher-level linguistic features such as morphology, lexis and syntax and the pragmatic applications of those units. The non-linguistic elements could include hesitations and other vocal sounds such as coughs.

Traditionally, a structured auditory phonetic analysis has been the leading way to analyse speech samples together with measuring a few acoustic parameters for comparison. Over the last decade, automatic systems for voice comparison have developed dramatically and have more and more become a major part of forensic speaker comparison. Therefore the interest in finding ways to compare "classical"/traditional parameters, human- and machine-based systems and calculating likelihoods under different hypotheses has increased.

This thesis involves three main objectives in five different studies. (For a full description see 5.1)

A. Robustness of the fundamental frequency (F0) as a parameter for FSC (study I).

B. Voice-similarity comparisons between perceptual judgments from different types of experiments and an automatic system (study II, III and IV).
C. Forensic comparison of speech in a case of disputed utterance (study V).

Objective A involves one study of the robustness of a new measure of one traditional phonetic parameter used in forensic speaker comparison: fundamental frequency. Objective B involves three connected studies where the first investigates the similarities and differences between pairwise perceptual judgments of voice similarity compared to results from an ear witness line-up experiment to establish to what extent it is possible to predict the outcome. The second study compares human perceptual judgments of voice similarity to the similarity calculations made by an automatic system. In the third study, a new way of comparing the performance of a human perceptual system and an automatic system is developed.

Forensic phonetics involves much more than comparing speakers; another major part is the transcription of recordings and in particular so-called cases of disputed utterance. This area is covered in objective C. As the word “forensic” implies more than actual comparisons in casework, a section has been devoted to clarifying some methodological issues in regard to the evaluation and presentation of forensic phonetic casework.

The introduction of this thesis starts with a detailed account of a very likely scenario for a forensic phonetic case in a Swedish setting. That section is followed by an extensive literature review and historical background to the subject of forensic speaker comparison. The objectives of the studies included in the thesis are then described and a summary of each paper is given.

4.1.1 A likely scenario for a forensic speaker-comparison case

The investigating police officer suspects drug traffic between two gangs of previously known criminals. She requests permission from a court to tap certain telephones since she suspects serious criminal activity. The court approves, considering the severe consequences of the suspected criminal activity. The telephone conversations following the tapping confirm the suspicions of the investigating officer. As a consequence, a member of a criminal gang is arrested and interviewed. The interview is recorded by the interrogating officer using a digital voice recorder, and stored digitally. The suspect denies being the person who is speaking in the recordings, where the officer claims that it is the suspect who is speaking. The suspect persists in denying the accusation, and as a consequence the recording of the interview, together with the recorded telephone conversations, are sent to the National Forensic Centre (NFC), with a request for a speaker-comparison analysis.

The recordings are then passed on to the analysts of forensic speech material for a quality check. The quality check, or so-called screening, will determine whether the recordings qualify for a full speaker-comparison analysis. For example, the recordings in question might be too short or contain too much background noise to qualify for a full analysis. If the recordings pass the quality check, a speaker comparison case is opened and the analysis starts.

The analysis consists of three independent parts. Part one involves editing and linguistic phonetic analysis of the recordings. This means that an analyst listens to the recordings and selects sections for use in parts two and three while transcribing and analysing the linguistic behaviour of the suspect and the recordings in question. The analyst can also prepare a so-called blind test for another analyst. In the blind test, a different analyst is presented with anonymised recordings from several speakers and is requested to compare each known speaker with unknown speakers and provide a conclusion for each comparison.

Part two consists of acoustic measurements of different vocal features such as fundamental frequency, speaking rate and vowel formants. Part three is then a so-called biometric voice-quality comparison using an automatic system. This system involves a trained background model of what general voices sound like (in some systems this is independent of language spoken, based on several thousand recordings of different voices). Acoustic features are extracted from the suspect and the recordings in question, and a voice model of the suspect's voice is created. The same acoustic features are then extracted from the recording in question and tested against the voice model of the suspect, and a likelihood score is calculated of how similar the test voice is to the suspect model. The score can then be normalised using a reference population model (recordings from a general population with similar recording features to the suspect model). It is also possible to use so-called impostors to normalise the test voice scores. Using the scores from the reference population and the actual score in the case can then provide a ratio between two hypotheses:

- *The test voice is the same as the suspect model.*

- *The test voice is not the same as the suspect model.*

The likelihood ratio will then tell us how much more likely it is to get the test score if it is the same voice compared to if the voices are different.

4.1.2 A likely scenario for a forensic disputed utterance case

In an authentic example of a different kind of case, the investigative officer records an interview with an eyewitness. In the interview, the officer perceives or notices the word "dom" (English "they") in the witness's statement describing the murder. This word becomes an important clue in the investigation. At a later stage, the witness claims not to have said "they" in that specific part of the interview, but instead a suspect's name "Tim". To clarify what was said the investigative officer requests a forensic disputed-utterance analysis from NFC. In situations like these, the NFC again sends the recordings for a quality check and if this is positive, an analysis is begun. In the case of a disputed utterance, the acoustics of the disputed part of the recording is analysed, and depending on the phonetic content, different acoustic features are extracted. In this case the same recording contained undisputed parts where the same or similar content was present and in such cases, those parts are analysed too as background models. It is possible that several instances of the disputed content must be recorded in a similar environment by other similar speakers and those recordings analysed in the same manner. The likelihood of the acoustic result is calculated under two different hypotheses, and the ratio between those shows which one is more likely given the result.

4.1.3 A likely scenario for an ear-witness line-up

In a similar case, a different witness heard the voice of one of the perpetrators and the investigative officer would like to test whether the witness can recognise the perpetrator if listening to a suspect's voice, which the officer believes to be the same. A request is again sent to the NFC and then forwarded to the forensic phonetic analyst.

In this case, recordings of the suspect and suitable foils have to be made. The witness should then undergo an ability test to see whether he or she can manage some more general ear-witness line-up tests before the witness is subjected to the real test following a set of criteria. These tests are normally very hard for a witness, the setup is very tedious to produce, and it is difficult to select and properly record the foils. However, when a witness has finally gone through a line-up test, the result is reported together with criteria and a setup description so it can be fairly judged by the court as valid or not.

4.2 General introduction to the included studies

This section provides a more general introduction in relation to the studies included in the thesis.

In forensic speaker comparison, there have been several different experimental and non-experimental approaches over the last few decades. The misconception about visual identification of voices (sometimes referred to as the voiceprint controversy) started as early as during the historical development of the spectrograph and the interpretation of both voice and speech through visual inspection. To introduce the reader to the subject a diachronic journey is needed to understand the development of some views and methods. A background section on the historical development (in 4.3 Historical Background) provides the reader with the tools to grasp later references.

The topic of forensic speaker comparison is a complex one. In other forensic analyses such as DNA analysis, there is a trace containing a segment that carries genetic information that can be connected to an individual with a certain level of precision that one can approximate and calculate. The sample can of course be difficult to extract depending on the quality of the finding, but the essential point here is that if one can extract this individual information one can calculate the ratio of the probability of the finding belonging to a suspect vs. the probability of the finding not belonging to the suspect but to someone else in a relevant reference population.

When it comes to a voice, a trace can be found on a recording of some kind, for example a bugged telephone conversation. The recording can be of different audio quality depending on encoding (compression, lossy or lossless, sample rate and bit depth) and background noise affecting the recording in different ways. This means that the main difference compared with other forensic analyses is the enormous variation in several dimensions a speech recording is affected by. There is (in the case of comparison) a reference recording, for example a police interview, with a certain audio quality. This audio quality is the first dimension of variation one is exposed to. To analyse the trace recording, one often (at some level) has to check for inconsistencies in the recording to make sure that there is only one speaker in the recording. The second dimension of variation is the intra-variation of the voice, which has two levels. In this thesis these two levels will be referred to as behaviouristic and biometric. The behaviouristic part is here the ways in which one speaks, which is presumed to be a learnt, inherent process affected by psychological patterns such as style, and sociophonetic, sociolinguistic and other situational factors. The biometric part is here referred to as the carrier of the behaviouristic part, i.e. the sound of the voice, affected by many layers of variation, but with a core, that is in this work presumed to be less varied i.e., a core that one attempts to capture in an analysis connected to the biologically constrained shape and form of a vocal tract and/or the vocal folds. Biometrics is used here simply in the same way as in many other definitions, i.e. to refer to metrics related to human characteristics. That means that there is no general exclusion of behaviouristic traits in the sense of the definition. The intention of the thesis is to show that it is extremely difficult to capture this core.

The first endeavour is to find a less variable and statistically measurable unit for the vocal folds' vibrations, i.e. fundamental frequency (F0), for two reasons. The first is that it is fairly easy to automatically collect and extract large amounts of data without manual labelling. The second is that one can then calculate how common or uncommon certain features are in that material if there is a relevant reference population. No matter how similar or dissimilar the samples in a case are, it will be impossible to tell how much more likely the results are if there is not a proper reference population to compare to. This is called distinctiveness or typicality. If a feature is measurable and easy to capture, that feature can be considered robust. A returning concept in the thesis is *robustness* in forensic phonetics, which will be defined in section 5.3.2 *The concept of robustness*. Two more concepts will also play a major role, as mentioned in section 4.1 General overview - voice and speaker. Voice is in the first study used in the narrowest sense, which is referring to the sounds produced by the vibrating vocal folds, but the definition is extended to include fricative sources and the influence of resonances in the vocal and nasal tracts, although the latter features are not treated to any great extent in the first study. Defined this way, voice comes very close to what is often referred to as *timbre*. Speaker, on the other hand, refers here to the linguistic, social, pragmatic, behavioural and idiosyncratic side of vocal communication, such as the use of the speech sounds of a given language, prosody, dialect or accent, idiolect, etc.

Most automatic voice comparison methods or speaker verification systems are based exclusively on vocal tract traits, although some systems have attempted to include speech factors. It is important to distinguish between these two parts as they are analysed in different ways and the results are difficult to combine. It is the intention of this thesis to clarify sub-parts of voice biometric parameters and speaker behaviouristic parameters and make an attempt to compare them in the best possible way for an optimal forensic phonetic analysis. The common denominator is a focus on *robustness* of parameters and the combination of the acoustically based measurements and the perceptually based analysis when comparing voice and speakers.

In addition to describing the different approaches and commenting on their usefulness, this work attempts to understand the different stages of development of this area of research and application. One strategy has been to use existing resources to the largest extent. That implies gathering data from existing databases and using them for analyses or finding open source software that can be used as they are or after being slightly altered. In the experimental work of the included papers, several tools and packages were used or developed. These tools and databases are described in section 5.4 .

Studies II, III and IV deal with the comparison and understanding of the processing of voice-similarity judgements by both human perception and an automatic voice-comparison (open-source) system. Paper two focuses mostly on finding similarities in how humans judge voice similarity using data from an ear-witness line-up experiment and then comparing the judged similarities to the actual ear-witness performance together with data from articulation rate (including pausing). Paper three instead compares the same judged voice similarities to the similarity scores from an automatic (open-source) voice-comparison system. In paper four, a new methodology on how to convert similarity judgments to scores to make better comparisons to automatic systems is presented.

The fifth paper deals with an actual forensic case of disputed utterance where the author acted as an expert witness. The data is then treated in a new way to explore methods for calculating an actual likelihood ratio using sparse data.

4.3 Historical Background

Diachronically, one can see forensic phonetics as something that has been an active part of legal systems through all times in one way or another. Recognising voices is described as far back as in the bible. “The voice is the voice of Jacob, but the hands are the hands of Esau” (Genesis Chapter 27, Verses 22-25; Alexander, 2005).

There is evidence that identifying our mother’s voice is a primary function of human aural perception at birth. It is possible that recognising your mother's voice was more important than using aural perception to understand language and communicate (DeCasper & Fifer, 2004). At least, it is obvious that even before birth we are under the influence of external auditory stimuli (DeCasper & Sigafoos, 1983; Spence & DeCasper, 1987) and we seem very early on to be able to discriminate between languages through speech rhythm (Nazzi, Bertoncini, & Mehler, 1998; Ramus, Hauser, Miller, Morris, & Mehler, 2000). For the thesis, it has become important to try and separate the inherent co-analysis of voice and speech. The way speech sounds are analysed seems to be closely connected to how listeners analyse voice, even as new-borns (DeCasper & Spence, 1986).

When his father, Isaac, recognises Jacob, this is a case of naive voice or speaker recognition. It is not possible to say whether Isaac recognised Jacob through his voice or through his speech, following the distinct definitions given in 4.1. Many experiments have shown how variable naive voice/speaker recognition is and how different listeners respond to different cues in different circumstances (Ramos, Franco-Pedroso, & Gonzalez-Rodriguez, 2011). An expert in FSC, on the other hand, is someone who is well educated on the different parameters used to describe voice and speech features and their variability in a structured manner (Schwarz et al., 2011a). In early history, i.e. before the development of modern legal systems, it is of course the former that is referred to.

Through history, voice and speech evidence, as with many other kinds of evidence, was considered reliable depending on how and who gave the testimony. One such example is the trial of William Hulet in 1660 (Eriksson, 2005). A witness had heard the face-covered voice of the executioner of King Charles I and declared that the speech was recognisable as that of Hulet, who was well known to the witness. Hulet was sentenced to death but later acquitted as the regular hangman confessed. This kind of misidentification was probably not uncommon at the time and probably also happens today. This is one historical example of one of the complex issues related to so-called naive speaker identification, namely to recognise familiar vs. unfamiliar voices. Maybe the most famous case more recently in history is the one involving Charles Lindbergh in 1932. It was a kidnapping case where Lindbergh, after written communication with the kidnappers, decided to pay the ransom for his son. He drove a negotiator to a cemetery with the ransom money and could only hear a kidnapper call the negotiator with the words “Here, Doctor. Over here! Over here!” Later the son was found dead. In 1934, a suspect, Bruno Hauptmann, was identified by Lindbergh almost 2.5 years after the incident at the cemetery. He later testified that he recognised Hauptmann’s voice in court. This brings up another difficult issue with naive speaker recognition, namely the influence of time delay and the memory of the ear witness.

Looking instead at speaker identification made by experts, one can say that it did not start until it was actually possible to record speech onto some kind of usable media in the early 1930s. Even then it was not very practical to carry around a recorder, but when telephones started to be used more frequently, crimes committed over that network of course became more common. One thought was that to be able to analyse recorded material, some kind of visualisation of the acoustics had to be made. The first and most important step in that development was the invention of the spectrograph. The major inventions were made at Bell Telephone Laboratories in the 1930s and the beginning of the following decade. Commercially it was sold under the name Sonagraph. The spectrograph was suitable for acoustic analysis as it could print energy in different frequency bands over time slices. Unfortunately not much was published on the new technology as it was classified as a war project until the end of World War II (Potter, 1945). The primary motive for the development was to advance phonetic research on speech and acoustic speech patterns. Another motivation was to develop a kind of sound-reading device for the deaf.

When the post-war development of forensic speaker comparison is discussed, which is the main interest here, it started with the development of the so-called voiceprint (later aural/spectrographic) identification method. This is the point of departure for our journey through the background development of forensic speaker comparison. The diachronic search through history will present work not directly tied to the issue of voiceprinting, but that relates to other aspects of the development of forensic speaker comparison that are equally or more important.

4.3.1 The Voiceprint Controversy

Over the years, going all the way back to the Second World War, speaker identification by spectrograms has had an influence on what most researchers today have agreed on calling forensic speaker comparison. The so-called voiceprint method has been criticised or embraced by different people at different points in time.

4.3.1.1 The early development and critics

It seems as if a common view in relation to the early stages in the development of voiceprint is that not many phoneticians contributed a view or opinion of the method (Hollien, 1977). However, many well-known scientists have written papers on the issue. Here, the proponents' arguments and credibility will be discussed.

Towards the end of the 1930s and the beginning of the 1940s, the sound spectrograph was (as mentioned in the last section) developed as a means of visualising the speech signal. During the Second World War, the work on the spectrograph was classified as a military project because the military saw the possibility of using the method as a way of identifying enemy troop movements from intercepted radio communications and telephone exchanges (Grey & Kopp, 1944; Meuwly, 2003a, 2003b). Potter (1945) (at Bell Labs) reported in his paper "Visible Patterns of Speech" about the new method and different ways of implementing the spectrographic technique in different applications for hearing-impaired people. The first academic paper on the subject of identification by voice is probably Pollack, Pickett and Sumby (1954). No visual examination of spectrograms was performed, but identification was done solely by ear and the general conclusion was that the duration of the speech signal was a particularly important factor in successful identification.

The focus on speaker identification decreased for a period of time immediately after the war, but attracted new interest when the New York Police Department started to receive reports of bomb threats to different airlines. They then turned to Bell Laboratories and asked if spectrograms could be used as a means of identifying the callers. Lawrence Kersta, an engineer at Bell, was assigned the task of investigating the matter (Owen & McDermott, 1996). After studying the matter for two years Kersta had become convinced that spectrograms could indeed be used to identify speakers. In the paper 'Voiceprint Identification' (Kersta, 1962a), he refers to the method as voiceprinting in direct analogy to the term fingerprinting. His paper only superficially describes the pattern-matching procedure, and instead focuses on the results from an experiment he conducted using high-school girls as a demonstration of the simplicity of the procedure. According to the identification results, the schoolgirls performed with remarkable accuracy (results presented in table summary). Kersta (1962b) presented a paper named “Voiceprint-identification infallibility” at the Sixty-Fourth Meeting of the Acoustical Society of America, where he refers to the earlier results. The emphasis in that paper was to show that the possible problem with disguised voices was handled very well by the visual inspection of spectrograms, even if skilled imitators performed the task. Several investigations followed, trying to establish how well one can actually identify a speaker using the method. Experiments on disguise were performed with differing results, most of them criticising Kersta and his method. Bricker and Pruzansky (1966) discovered that it was more difficult to compare samples with /i/ than /a/ and that context dependence is important in order to be able to perform the identification. They also suggested that perceptual analysis might enhance the identification rates. Young and Campbell (1967) tested the effect of different contexts and made a summary of the experiments that had been performed to date (see Table 1).

Table 1. Results of early speaker-identification experiments (Young & Campbell, 1967).

Experimenters	Pollack et al. (1954)	Kersta (1962a)	Bricker & Pruzansky (1966)	Young & Campbell (1967)
Results: correct identifications (%)	84-92	99-100	81-87/89	37.3-78.4
Method	Short words, isolation	Short words, isolation and context	Words, isolation and context	Short words, isolation and contexts

The differences between studies were substantial. However, the results are difficult to compare due to the lack of documentation and differences in methodology. The suggestion from Bricker and Pruzansky (1966) to use spectrograms (i.e. voiceprints) in combination with a perceptual approach was tested the following year (Stevens, Williams, Carbonell & Woods, 1968). It was discovered that in this test, subjects had an error rate of 6% with a perceptual approach and 21% errors using solely visual spectrogram pattern matching. This was one of the first clear indications that including perceptual aural examination is crucial. In the case U. S. v. Frye, 1923, where an early version of a so-called lie detector test was presented as evidence, the court dismissed the evidence saying that for an expert testimony to be admissible, the method on which it is based must be “sufficiently established to have gained general acceptance in the particular field in which it belongs”. This principle has not been applied in all states and the interpretation of exactly what it means for a method to “gain general acceptance” has often been disputed, but the ruling has nevertheless often been used as a motivation for not accepting voiceprint testimonies (Tiersma & Solan, 2002; Keierleber & Bohan, 2005).

4.3.1.2 Discussion of the methodology

The first attempts at introducing voiceprint testimonies were all dismissed with reference to the Frye ruling (Owen & McDermott, 1996). This discussion is similar to the one that had been going on in Europe, where some French scientists for quite some time had stated that any kind of speaker identification made by experts would be unethical (Chollet, 1991). Others claimed that it would be unethical *not* to do it, considering the unchallenged testimonies by lay experts (Boë, 2000; Braun & Künzel, 1998). Opinions quickly became divided after the introduction of the voiceprint technique. Proponents (some scientists, but mostly laymen) defended the technique, regarded it as highly reliable, and appeared as expert witnesses in various criminal cases. Most scientists, however, were sceptical, regarding it as not sufficiently tested (e.g. Stevens) or dismissed it completely (e.g. Hollien). Bolt et al. (1969; 1973) criticised the method and presented several relevant questions:

- When two spectrograms look alike, do the similarities mean that the speaker is the same or merely that the same word is spoken?
- Are the irrelevant similarities likely to mislead a lay jury?
- How permanent are voice patterns?
- How distinctive are they to the individual?
- Can they be successfully disguised or faked?

Expert witnesses did not agree as to its reliability, and various courts of law have ruled both for and against the admittance of such evidence. One response declared “It is our contention that opinions based on feelings other than in actual experience are of little value irrespective of the scientific authority of those who produce such an opinion” (Black et al., 1973). However, this response did not contain any scientific evidence supporting the method. In the midst of this heated debate the IAVI (International Association of Voice Identification, later called VIAAS, Voice Identification and Acoustic Analysis Subcommittee) was founded (1971) on the initiative of Kersta, presenting guidelines for the practitioners of the so-called aural/spectrographic identification method. The IAVI stipulated that one needs two years’ apprenticeship supervised by an authorised examiner. Five levels of identification were used as alternative decisions (Owen & McDermott, 1996):

- Positive identification/elimination
- Probable identification/elimination
- No opinion

The critique and criticism continued during this period, presenting results showing change as a function of age as well as disguise by imitators (Endres et al. 1971) and emotional states (Williams & Stevens, 1972). Other researchers, however, published results that seemed to lend support to Kersta’s method. A group of researchers led by Oscar Tosi at Michigan State University tested Kersta’s methods in an extensive study that produced results which very closely matched Kersta’s: 6% false identification and 13% false elimination. If all ‘uncertain’ responses were excluded, there were only 2% false identification and 5% false elimination, thus supporting the “no opinion” criterion given by the IAVI (Tosi et al., 1972). Criticism followed, stating that there was an exaggeration in the interpretation of results: “The data show that the system tested does not effectively reduce the effects of contextual variation, and cannot be used for either absolute identification, elimination, or population reduction” (Hazen, 1973). Hollien (1974) commented on the dispute, also referring to the “social relevance of the problem”, meaning there was an ethical problem which was not being regarded in the discussion about using the method or not. While one should try to ensure that justice is done as much as is possible, one cannot go so far as to use unreliable methods that are not supported by the greater part of the relevant scientific community. Several papers from IAVI members followed that supported the method. Hall (1975) stated, “Variability does not exist”. Smrkovski (1975) published a paper on the importance of experts performing the aural-visual identification. In the paper, examinations performed by trainees and “professionals” (minimum of two years of field experience) differed slightly. Trainees produced 0% false identifications and 5% false elimination and responded “no decision” in 25% of the cases, while the “professionals’” results were 0% false identification, 0% false elimination and 22% “no decision”, thus showing the relevance of field experience. The proponents’ activity had increased as a response to the earlier criticism and it was time for a new response. One of the most critical scientists who was active in the field at this point was Harry Hollien. Hollien and McGlone (1976) tested the method on five faculty members and one graduate student. They performed visual comparisons among twenty-five faculty members and graduate students at the University of California. The results concluded that “/.../ even skilled auditors such as these were unable to match correctly the disguised speech to the reference (normal) samples as much as 25% of the time /..../” (Hollien & McGlone, 1976). In a similar study, Reich et al. (1976) studied effects on six vocal disguises. Four trained spectrographic examiners achieved 56.67% accuracy in matching the undisguised material. The authors concluded “The inclusion of disguised speech samples in the matching tasks significantly interfered with speaker-identification performance”. Hollien, (1977) also published a “status report” explaining the uninterrupted use of the method by making four claims:

- Proponents of voiceprints are rarely opposed.
- They claim their method meets the Frye test.
- They claim uniqueness.
- They claim that their research demonstrates reliability and that their voiceprint examiners can correctly use this tool.

Other studies in the same year included Houlihan (1977a, 1977b), two tests by twenty-one undergraduates making visual comparisons of disguised speech from a homogeneous group of nine women and five men. Correct identification for normal speech was high (95-100% correct), a bit lower for lowered pitch (85-95% correct), falsetto (90-95%), muffled (75-100%) and whispered as wide results as 5-98%. She interprets these results as being supportive of voiceprint identification of normal speech. Rothman (1977) further concluded, “Aural method is clearly superior to the spectrographic or ‘voiceprint method’”. Michigan State University also presented several interesting papers regarding the method, all supporting it.

4.3.1.2.1 Oscar Tosi

The ox pulling the scientific carriage in the 1970s was Dr. Oscar Tosi. At an early stage of developing the method he had testified against the use of spectrographic comparison, but in the case Trimble v. Heldman (1971), the Supreme Court held that ”spectrograms ought to be admissible at least for the purpose of corroborating opinions as to identification by means of ear alone”. Tosi had impressed the court, claiming high reliability of the technique after testing it (Owen & McDermott, 1996). Tosi and Greenwald (1978) presented an experiment at the sixth meeting of the IAVI, including aural, visual and combined methods for identification by professionals and trainees (only two weeks of training). The material used came from a minority group (described in the study as twenty-five male and twenty-five female Chicanos). The number of trials per examiner was as many as 600, and the results were 0% errors by the professionals using combined aural-spectrographic identification and, in contrast to earlier studies, professionals used “no opinion” more frequently than the trainees. The following year, Greenwald (1979) presented his master's thesis testing effects on decreased frequency bandwidths in aural-spectrographic examination. The examiners were again professionals (three examiners with more than eight years’ experience) and trainees (five examiners with less than two years of experience). The professionals again produced 0% errors whereas the amount of “no opinion” increased with restricted bandwidths. The trainees’ results were not much less comforting, with an average of 6.1% false identification and 4.1% false elimination at restricted bandwidths, but 0% errors at the greatest bandwidth tested. Later Tosi published a book (Tosi, 1979) where he gives an up-to-date reference to all subjects involved in forensic phonetics. It also criticises the authors mentioned earlier who opposed aural-spectrographic examination. Even though the book gives a complete picture of speech acoustics and its reflection in spectrographic representations, his conclusions are far too wide and are based on a few small-scale experiments and a methodology that clearly lacks validity due to the variation found in the visual representations of speech. Along with his colleagues at Michigan State University, he was one of the founders of the Forensic Science program. Thanks to him, the term “voiceprint” was excluded as a term as he, in opposition to Kersta, did not propose the method’s infallibility. It was probably because of him that the IAVI was absorbed into the IAI (International Association for Identification) in 1980.

4.3.1.3 Modern history from 1980s until the millennium shift

A period of status quo followed until it was revealed in 1986 that the FBI was using the method. By this time, it had been used for investigative purposes since the 1950s (Koenig, 1986b). Koenig (1986a) reported error rates for the spectrographic voice-identification technique under forensic conditions, stating, “The survey revealed that decisions were made in 34.8% of the comparisons with a 0.31% false identification error rate and a 0.53% false elimination error rate.” The report/survey was rather limited in its explanation of the figures though. In a reply, Shipp et al. (1987) presented relevant criticism of the methodology.

- What procedures do they (the FBI) actually use when employing the method?

The results were based solely on the feedback from verdicts, which was in a sense circular, since the technique employed might determine guilt or innocence and that verdict then was used to verify the results. The extended answer to this reply cleared up some of the confusion in the first paper, but reported no more evidence as to why the method was employed at all. According to the FBI survey, voice-identification examiners at the FBI had to have a minimum of two years of experience and to have completed at least a hundred voice-comparison cases. Combined aural-visual examination was employed and decisions used were: very similar, very dissimilar, no decision (low confidence) (Koenig, 1986b). Further, they had to have “/.../ a minimum of a Bachelor of Science degree in a basic scientific field, completed a two-week course in spectrographic analysis” (“/.../ or equivalent”) and pass a yearly hearing test. The VCS (Voice Comparison Standards) of the VIAAS (Voice Identification and Acoustic Analysis Subcommittee, 1991) are very similar and obviously not independent. The criteria of the VIAAS included a high-school diploma instead of a bachelor’s degree, a ten-word comparison vs. twenty and they did not require the recording to be an original. In Koenig (1986a), it was more or less just stated that the recordings from suspects should in some way be similar to the reference material, i.e. “/.../ a spectral pattern comparison between the two voice samples by comparing beginning, mean and end formant frequency, formant shaping, pitch timing etc., of each individual word”, which does not clarify the question of why or how. In the VCS, there are at least some general descriptions of what to look for in the visual comparison such as “/.../ general formant shaping, and positioning, pitch striations, energy distribution, word length, coupling (how the first and second formant are tied to each other) and a number of other features such as plosives, fricatives and inter formant features”. The number of alternative decisions was seven and included identification, probable and possible identification as well as elimination (Gruber & Poza, 1995). The greatest issue today is perhaps the common opinion expressed by media that “The CIA, FBI and National Security Agency have computers that use special programs to identify voiceprints. The idea is that every voice has a unique pattern like a fingerprint.” (CNN website, December 2002, in conjunction with the Bin Laden voice affair) (Rose, 2002).

The use of aural-spectrographic voice-identification evidence can still be found, usually performed by private practitioners who have no special skills, but unfortunately also in national forensic laboratories in different parts of the world (Morrison et al., 2016). At least as recently as 2006 the FBI was still using it for investigative purposes, and so were the Japanese police according to Rose (2002). In 2002, it was still admitted as evidence in some states in the US, and at least one case involved voiceprint evidence in Australia in 2002 (Rose, 2003).

4.3.1.4 Other summaries

Nolan (1983) contains a thorough summary up to the 1980s, with very relevant comments such as:

> “/.../ the voiceprint procedure can at best complement aural identification, perhaps by highlighting acoustic features to which the ear is insensitive; and at worst it is artifice to give a spurious aura of ‘scientific’ authority to judgements which the layman is better able to make.”

Chapter Ten in Hollien (1990) gives a more recent summary (up to the 1990s) and provides an insight into the voiceprint cases he has been involved in. Künzel (1994) gives a European perspective as well as constructive criticism. Gruber and Poza (1995) give even more background, devoting a whole chapter to the issue and providing some ”inside information”, as the second author “/.../ was technical adviser for an important monograph commissioned by the FBI in 1976 to evaluate the method” and had also completed the two-week course given by Kersta. Owen and McDermott (1996) contain a well-organised summary of all tests conducted with the method, including all relevant information from each paper. Broeders (2001) provides an overview of forensic phonetics, stating which methods are used all around the world and by which people in which country. Rose (2003) summarises everything up to the millennium shift and is most valuable because of its comments on different issues. The Frye test basically concludes that new scientific evidence should have gained “general acceptance” in the relevant scientific field. In 1993, the Daubert case set a new standard of interpretation now accepted by several American courts of law, stating “good grounds” in validating an expert’s testimony. Aural/spectrographic/voiceprint identification has several methodological problems that have not been dealt with in the literature.

- What is it that one is supposed to look for?
- What signals identification?
- When are spectrograms similar enough to indicate the same speaker?
- When are they dissimilar enough?

Even though several experiments have shown that spectrograms are not reliable in verifying identity, none of the papers conclude how representative they are of a speaker’s voice.

- Can one make a reliable decision using spectrograms?
- Finally, has the method ever been one that is accepted by the relevant research community?

Generally, a majority of the relevant scientific field knew that the most reliable way of comparing voices at that time was by aural perceptual analysis by a trained phonetician, but many researchers still remained passive (Hecker, 1971).

4.3.2 Forensic phonetics in Sweden

Almost all forensic analyses in Sweden are handled by the Swedish National Laboratory of Forensic Science (Statens Kriminaltekniska Laboratorium, SKL) renamed in a huge reorganisation of the Swedish police as the National Forensic Centre (NFC) from January 1st 2015. Until 1994, there were no regular forensic phonetic analyses being performed at the lab, even though occasional cases were handled by external academics. In 1995, SKL employed a full-time phonetician to work on forensic speaker-comparison cases. He was employed for approximately 10 years (2006) and the lab performed approximately 30-35 forensic speaker-comparison cases per year. Due to other priorities, no one was employed after that (except short-term contracts) and the laboratory more or less stopped performing the analyses in-house.

Several independent police districts started requesting this kind of forensic analysis in 2005. During the course of this research a new independent laboratory was started, and the cooperation with NFC (formerly SKL) has increased over the years. Since 2011, the company Voxalys AB (Aktiebolag) has been a subcontractor to the NFC. In total, the lab handles between 30-40 cases per year and most of them are speaker-comparison cases.

4.4 Short Review of Forensic Phonetics

Reviewing the field, there are mainly two branches simultaneously working on one goal. One could be called the phonetic approach and the other an engineering approach. Traditionally, it is not only the use of an automatic method itself that has been the main difference between the two. A divider has been the discussion on sampling, reference data and statistical testing to verify the reliability of FSC (Boë, 2000; Hughes, 2014), in large part driven by the needs of the different methodologies. Generally, this means that automatic methods use much larger quantities of recordings for analysis of voice, but with a more holistic approach. By contrast, the phonetic approach typically involves much more detailed analysis of smaller corpora to capture the many behaviouristic dimensions of individual speakers.

When the everyday life and kind of casework of the forensic phonetic analyst became more obvious to those working in the engineering world, i.e. until recent years predominantly short, noisy recordings in mismatched conditions, they became aware of the more general problems of using an automatic system. However, some attempts and co-operations have been going on between phoneticians and engineers (e.g. Magrin-Chagnolleau, Bonastre & Bimbot, 1995; Schwarz et al., 2011b; Hughes, Wood & Foulkes, 2016).

4.4.1 A classical phonetician's approach

The phoneticians' approach has been the leading one in forensic speaker-comparison casework in Western Europe and Australia (Foulkes & French, 2012; Rose, 2002). If the development of forensic speaker comparison is analysed diachronically, one could say that the 'voiceprint approach' was invented by a phonetically rather naive engineer. By, more or less successfully, diminishing the number of "voiceprinters", the classic phonetic research approaches took over. That at first meant mainly perceptual analysis. Most phoneticians come from a linguistic research environment, which gives a very wide knowledge range and background for performing analysis of speakers. Often they ignore the fact that it is the comparison of speakers that is done and not just an analysis of the similarities in voice. By that it is implied that there are many idiosyncratic features in speaking that are introduced through behaviour, that specific behaviour might influence what kind of language is used in a certain social situation, thus the forensic phonetician might make reference to aspects of phonology, syntax, pragmatics etc. as well as acoustic phonetic features (Foulkes & French, 2012).

The forensic phonetician typically makes use of small corpora of speakers with detailed analysis of multiple linguistic components, supported with reference to published literature e.g. to estimate the distribution of a given feature in different dialects. It is also possible to gather and/or analyse recordings for comparison to the evidential recordings in a given case. In recent years the analysis of recorded voices has become more widespread in casework because advances in technology mean it is easier to collect and analyse recordings for reference. Historically, this was not the case. In the early development of forensic phonetic techniques, a set of ten or so speakers was a rather common data set (Atal, 1972; Compton, 1963; Glenn & Kleiner, 1968; Hargreaves & Starkweather, 1963). Recording and acoustic analysis were most certainly difficult until the 1990s.

4.4.2 An engineer's approach

In the same way that the classic phonetic approach could be considered technically naive, the engineer's approach is no less phonetically and linguistically naive. The very first step where one has to be careful is to realise that Kersta (1962b) and the voiceprint method appeared through an early naive engineering approach. He was not aware of the full extent of variation in speech and language behaviour. There was little attempt to investigate the relative performance of different phonetic features in distinguishing from one another, or to develop theoretical models of the voice or individual speaker behaviour.

Current AVC methods are very data driven and statistically based. However, this does not mean that there are no other approaches using more linguistic and phonetic knowledge that perform better, especially on difficult test samples (Schwartz et al., 2011a). With the evolving and emerging world of computer science, web science and e-science, databases and computer-based methods are being used by a wide range of scientists. One could therefore say that all sciences use engineering in one way or another. It is a slow process to learn how to use it better and how to standardise data to make research more effective and recyclable. Attempts are being made to improve the standards and usability of data and methods through projects like CLARIN[5] and META[6]. The pursuit of system improvement is both a downside and an upside for engineers. The downside is the vast number of statistical algorithms applied with minimal improvement. Many research teams develop a system of their own and there is no common baseline, i.e. the same algorithms might generate both positive and negative results. This could depend on bugs in the code, incorrect implementation of algorithms or just different effectiveness in the coding. Another common difference is the training data for background models, the main issue being that it is very hard to track back why a certain method seemed successful in one attempt and less successful in another test when two different systems are being used, even if they are well explained and seem to be similarly built. The upside is that it has made a very large number of researchers compete to become one of the best sites in evaluation campaigns such as the most common one arranged by the National Institute of Standards and Technology (NIST). Through the evaluations, a kind of baseline performance has been achieved by many systems (i.e. the difference between the best results is minimal), as well as new ways to improve them regarding mismatched conditions (Alexander, Botti & Drygajlo, 2004; Matrouf, Scheffer, Fauve & Bonastre, 2007). Even the inclusion of phonetics and phonology in systems has shown improvement (Reynolds et al., 2003). This means that it is now rather well known what one can expect as a baseline performance by a system using well-known common techniques in the kind of conditions in which they were tested. Another upside is that the development has been a positive injection regarding the statistical view on variation in forensic science and here especially within forensic speaker comparison.

If there is data that is needed to be able to train a biometric system, a further question is how to interpret the produced likelihood ratio? A suggested middle path for giving evidence using a Bayesian approach has been proposed (Champod & Meuwly, 2000; Drygajlo, Meuwly & Alexander, 2003; Meuwly & Drygajlo, 2001). These suggestions have sparked a discussion and generated new thoughts and ideas in the active forensic phonetic community. Several British forensic phoneticians accepted a proposal to adjust the impressionistic likelihood expressions used by experts (French & Harrison, 2007). However, there were disputes about how to interpret and present the results (Rose & Morrison, 2009; French, Nolan, Foulkes, Harrison &

[5] http://www.clarin.eu/ (Fry & Thelwall, 2008)
[6] http://www.meta-net.eu/ (Borin, Brandt, Edlund, Lindh, & Parkvall, 2012)

McDougall, 2010). Recently some British practitioners have adopted a similar approach to the one suggested by the European Network of Forensic Science Institutes (ENFSI) (Willis et al., 2015).

4.4.3 A possible new generation

Mainly through the inclusion of insights from linguistics and phonetics in language technology, a new, broader and maybe more complex generation has emerged (William & Barry, 2005). These junior researchers could have a background in either phonetics, linguistics or in computer science with an interest in language technology in general. In Sweden, this research area generated a National Graduate School of Language Technology (GSLT)[7] in 2001. Students with different backgrounds and from different disciplines came together and were required to study methods and problems occurring in other fields. This kind of effort has produced cross-disciplinary researchers who should not be defined in any classical narrow sense as either linguists or computer scientists. Their often broader backgrounds have made them competent to use methods and approaches from different fields related to language technology (LT). If you compare this to the two different approaches to forensic phonetics described here, this kind of background is very suitable. A natural way forward has been to bring the two approaches together in a logical and systematic manner. This means that one has to start investigating the overlaps and dependencies that are present between the robustness with which humans' judge voice similarity compared to an automatic voice-comparison system. That is one of the main aims of this thesis. The text-independent Automatic Voice Comparison system is here defined as a UBM-GMM (Universal Background Model - Gaussian Mixture Models, also referred to in the literature as GMM-UBM) system as described in Reynolds et al. (2000). The specific system used for the different experiments and tests in this thesis is described more closely in Bonastre et al. (2005). As the majority of forensic phonetic investigators are phoneticians, and have been so for a long time, there has been a kind of scepticism where the engineering community is regarded as not understanding real forensic casework problems. Some engineers have even suggested that experts should not have any opinion in forensic speaker-comparison cases as it was not possible at that time to quantify the reliability of the evidence (Boë, Bimbot, Bonastre & Dupont, 1999). This is a general problem for many different areas within forensics. However, the forensic phonetic community argued that all evidence is important, but that how the evidence is

[7] http://www.gslt.hum.gu.se/

evaluated in court depends on how it is presented, as well as on the expert's experience and knowledge (Braun & Künzel, 1998). There is also a risk that other non-experts will make judgements and present confident reports on issues they do not have sufficient knowledge about. This means it would be unethical not to help the cause of justice when this kind of evidence is at hand. Expertise in forensic science demands knowledge of many different fields and of forensic speaker comparison in particular.

4.5 Modern Forensic Speaker Comparison

Leaving all descriptions of voiceprints behind, the journey of exploring forensic speaker comparison as it works today can start. What is being done today and how? In 1991, a few years after beginning a series of conferences on forensic phonetics (1988-1990) in York, UK, the International Association of Forensic Phonetics (IAFP) was founded. Several well-known phoneticians who were interested in this applied area of their research were present. Up to this point, the field seemed to have been made up of several academics who were heavily criticised for their work. The association was started as a way to invite the critics into a kind of controlled environment for people who were active in the field. Even today the aim of the association is to:

A. Foster research and provide a forum for the interchange of ideas and information on practice, development and research in forensic phonetics and acoustics, and
B. Set down and enforce standards of professional conduct and procedure for those involved in forensic phonetic and acoustic casework (Sept 1st, 2009).[8]

Even though the association was set up and the first steps were taken towards a more unified practice, people today still work in rather separate fields and frameworks when doing forensic phonetic casework. Not until the 2003 annual conference in Vienna was the A for Acoustics added to the association's name. This was intended as an invitation for academics and researchers from engineering to join in. While research and methods on aural and "classic" acoustical analysis had carried on, the engineers, more focused on the domain of speaker verification for applications, had had their own developments for years through the National Institute of Standards and Technology (NIST) evaluations. NIST evaluations, started in 1996, were originally an annual event for institutes, commercial companies and academic research

[8] http://www.iafpa.net/

groups to test their automatic speaker-verification systems on the same data in order to be able to evaluate the performance of different methods. One of the problems with that is that all systems are developed with completely different computational systems in different labs, and the difference in performance cannot always be explained on the basis of the algorithms and statistics used, but might be influenced by the architecture of the system itself. Generally, these two different areas of research in recognising voices/speakers in some sense try to solve the same problem using different methods. More recently, especially in Europe, there has been a tendency to attempt to accredit all forensic analyses. This has generated projects and research on how the process should be carried out. However, there seems to be a lot of variation between countries in how these criteria are implemented (Drygajlo et al., 2015; Meuwly, Ramos & Haraksim, 2016).

4.5.1 Short review of acoustic measurements and automatic systems

After the addition of Acoustics to the name IAFPA, influence and participation from engineers has been more frequent at the annual meetings. At the same time, the performance of state-of-the-art automatic speaker-verification systems has improved substantially with new techniques emerging (Dehak, Kenny, Dehak, Dumouchel & Ouellet, 2011; Lei, Scheffer, Ferrer & McLaren, 2014). One could say that even before the organisation was founded, some early studies explored the value of different acoustic patterns for voice classification, for example, long-term spectral variation (Furui, 1978; Harmegnies & Landercy, 1988; Wendler, Doherty & Hollien 1980), which has continued to some extent (Lindh, 2005). With the development of technology, it became easier to measure acoustic parameters. Experiments were performed which tested other generally known phonetic acoustic parameters such as fundamental frequency (F0) (de Pinto & Hollien, 1982; Graddol & Swann, 1983; Horii & Ryan, 1981; Künzel, 1989; Künzel, Köster & Masthoff, 1995; Linville, 1988; Linville & Korabic, 1987; Stoicheff, 1981). Problems regarding the same parameter were later discovered (Braun, 1995; Boss, 1996; Künzel, 2000). Other measures have been suggested (Lindh & Eriksson, 2007) as the integration of the F0 parameter into automatic systems (Jiang, 1996; Reynolds et al., 2003). However the attempts to integrate fundamental frequency into systems had started much earlier (Atal, 1972). Last but not least, when looking back at the literature on acoustic measurements for speaker comparison, there is a considerable amount of published work on formant frequencies and formant contours (Alderman, 2004; Clermont & Zetterholm, 2006; Ingram, Prandolini & Ong, 1996; Kinoshita, 2001; McDougall, 2004, 2006; Nolan & Grigoras, 2005; Rose, 1999; Rose, Osanai & Kinoshita,

2003) and the problems with measuring them (Byrne & Foulkes, 2004; Imaizumi & Kiritani, 1989; Künzel, 2001, 2002; Moosmüller, 2001; Nolan, 2002b). Several attempts regarding the possibility of profiling body size as a function of acoustic parameters were made (Dommelen, 1993; Gonzalez, 2004; Greisbach, 1999). The inclusion of the A in IAFPA also caused groups of engineers to become more and more interested in the forensic domain of speaker identification. At the Technical University of Madrid, one group began to develop a very user-friendly system for forensic phonetic experts. At the Laboratoire Informatique in Avignon, one group started to develop an open-source system first called ALIZE (and for a period Mistral). An open-source system can play an important part in the future development and comparison of system algorithms in evaluations.

More recent developments include systems that attempt to include phonetic acoustic measurements within the frame of an automatic system. This has had some success and also introduced more straightforward ways to evaluate different measures in the same ways as automatic systems (Becker, Jessen & Grigoras, 2008; Jessen, Enzinger & Jessen, 2013). The state of the art has also in the last few years moved from the UBM-GMM systems (Reynolds et al., 2000) to i-vector based automatic systems (Bousquet, Matrouf & Bonastre, 2011), which reduce the dimensionality of the statistical modelling enormously and improve performance, especially on mismatched recording conditions.

4.5.2 Summary

This section has reviewed the historical development of forensic phonetics in general. History has shown great potential through the possible gain in collaboration between linguistics, phonetics and automatic speaker recognition research. The progress of FSC lies in this collaboration and the emergence of a new generation of researchers that have one leg in both disciplines and who promote collaboration. This thesis is a step in that direction.

5 RESEARCH FOR THE THESIS

5.1 Objectives

The overall aim is to seek ways to compare, contrast and ultimately integrate knowledge from linguistics, phonetics and automatic systems into forensic phonetics.

The three main research objectives are summarised as follows.

A. Study I - The foundation of voicing is vibration of the vocal folds. Many forensic phonetic studies have been carried out to investigate fundamental frequency (F0). The first objective has been to find the most appropriate measure of F0 and test the robustness compared to classical measures of F0 such as mean and standard deviation, and to investigate the discriminative power of the parameter among a set of voices.

B. Automatic systems for comparing voices have become more widely used for forensic voice analysis in recent years. However, the combination of these automatic systems (mainly research by engineers/computer scientists) with the more classical phonetic analyses (perceptual analysis and more manual acoustic measurements such as formants, F0 and articulation rate) have largely developed independently. Only recently they have begun to converge. The second objective is to find ways to compare human perceptual judgments of voice similarity of two kinds to the comparisons made by an automatic system.
    a. Study II - An investigation of how listeners discriminate and judge voice similarities in a pairwise experiment compared to the results of a simulated ear witness line-up.
    b. Study III - A comparison is made between human voice-similarity comparisons and the same comparisons made by an automatic voice-comparison system.
    c. Study IV - A new method is developed for converting human perceptual judgments into scores to create a better way to compare the performance of human judgments of voice similarity with that of automatic systems.

C. Study V - An attempt is made to calculate a likelihood ratio using sparse data in a forensic disputed-utterance case, using an example from real forensic case data.

5.2 Research questions

After defining the objectives of the thesis this section will describe specific questions regarding each objective.

5.2.1 Fundamental frequency from an individual perspective

One of the most frequently studied parameters is fundamental frequency, maybe because it is generally easy to measure even in a noisy environment. Some attempts have been made to quantify, i.e. assess the between- vs. within-speaker variation of, the data (Boss, 1996; Graddol & Swann, 1983; Jassem, Steffen-Batog & Czajka, 1973; Jiang, 1996; Künzel, 1989; Künzel, 2000; Lindh, 2006; Linville, 1988; Linville & Korabic, 1987), but in most cases without a forensic perspective. Braun (1995) gives a very thorough categorisation of various factors affecting F0 and summarises the findings of numerous studies. *Robustness* is a key feature for forensic speaker-comparison parameters. F0 seems to fulfil several criteria.

RQ1. What could actually cause the sometimes substantial intra-variation for fundamental frequency and how would it be possible to model it or ignore the redundant variation?

RQ2. Which long-term measure of fundamental frequency would be the most robust and usable in the forensic context?

5.2.2 Human perception and Automatic Voice Comparison from a forensic perspective

To fulfil the second objective, and to be able to compare human perception to comparisons made by a machine, an Automatic Voice Comparison (AVC) system is used. In order to fulfil requirements of reproducibility, affordability and in order to use well-known general techniques, an open-source system was used (i.e. a system that is licensed, with editable, adjustable and downloadable code). Key questions in this respect include:

RQ3. What can be predicted by comparing voice similarity for listeners in an ear-witness line-up compared to listeners' voice similarity judgments?

RQ4. Do the same mistakes occur with automatic systems as with listeners' judgments of voice similarity, and why or why not?

RQ5. Is it possible to compare the results of listeners' judgments of voice similarity and scores from an automatic voice comparison system?

To answer these questions, one first has to investigate the available background information on how automatic systems perform comparison between voices relative to human listeners. It has, for example, been shown that listeners seem to be more sensitive to speech features than voice characteristics, and have difficulties separating speech and voice when performing similarity judgements (Petrini & Tagliapietra, 2008). In contrast, AVC systems ignore speech features and claim to model purely the voice characteristics. This probably means that human and machine similarity judgements between a homogeneous group of speakers are done differently. We can therefore add the following question:

RQ6. Is the AVC system better at discriminating between voices than human listeners when speech feature similarities are in conflict with voice similarities?

An interesting example of how to try and understand this main difference in how systems and humans would compare speakers could be as follows: A pair of identical twins' voices might sound extremely similar. Their voice quality does not differ much, mainly due to the similarities in the structure of their vocal organs. Maybe their way of speaking is also rather similar as they have grown up together. However, one of the twins has a clear stutter. A hypothesis would then be that, ignoring the difference in the way they speak, the system would consider their voices to be extremely similar if even different at all. A listener would then instead consider the speakers to be very different due to the stutter that one of the twins has. With a better understanding of the differences and similarities of, on the one hand, systems analysing voice quality similarities and, on the other hand, analytical listeners analysing qualities of speakers' voices, it will be easier to combine the two in forensic casework. It is becoming more and more common to combine the methods in forensic phonetic casework, but it is unclear how we can combine them in the best way.

5.2.3 Forensic transcription and disputed utterances

Forensic phonetics does not only involve speaker comparison. Quite a few cases are concerned with transcription of recordings of different levels of difficulty. When it comes to general problems with forensic transcriptions, a few studies have been performed such as Fraser, (2003), which mainly is concerned with the background of the transcriber and in particular their background in linguistics and phonetics, which of course is crucial to most analysts involved in this kind of research and analysis. More recently, some very important questions have been raised regarding the use of transcriptions in court and the problems of contextual priming (Fraser & Stevenson, 2014). The issue of priming is essential in an auditory phonetic approach. If only limited work has been done on the combination of auditory and automatic methods in comparing voices and speakers, even less work has been done on combining automatic speech recognition (ASR) and forensic phonetic transcription (Lindh, 2008). Sometimes parts of a recording are disputed. These disputed parts can be treated in different ways, as there can be well-defined hypotheses regarding what is said. If this is the case, acoustic analysis of the recordings can add substantially to avoiding priming in the analysis process. Regarding the third objective one may ask:

> RQ7. How can a likelihood ratio be calculated for acoustic measurements from disputed utterances in a forensic analysis with very little reference data available?

5.3 Various general methods and definitions

5.3.1 Robustness of humans and machines

This section serves to familiarise the reader with two things. The purpose of the first part is to explain and give the reader a background to the complexities and choices made regarding the concept of robustness in the forensic phonetic setting of the thesis. This is done by dedicating parts of this section to describing the use of the concept itself and how it has been applied in forensic phonetic research. The most common parameters used in forensic phonetic analyses will be presented, with a particular focus on fundamental frequency. The second part gives an introduction to how automatic systems generally work and presents the open source framework used in different ways for experiments that are presented in the thesis.

5.3.2 The concept of robustness

The concept of robustness refers mainly to resistance to noise, duration and mismatched recording conditions in the technical sense (external), but also resistance to emotional variation and liveliness levels in the behaviouristic sense (internal).

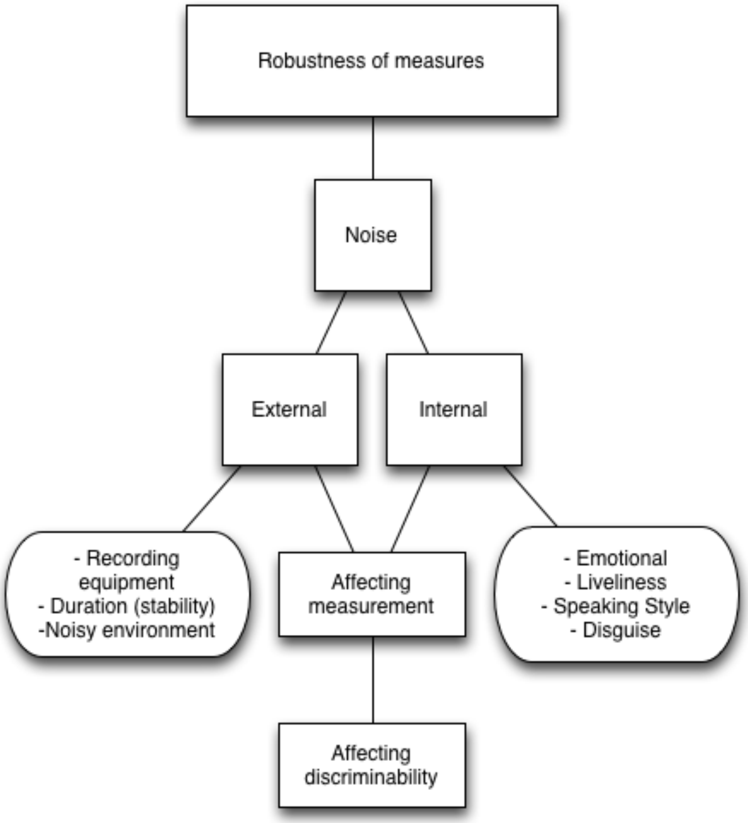


Figure 1. Robustness flowchart.

Figure 1 presents an overview of how robustness in conjunction with the resistance to noise can be defined. Robustness of a measure can here be defined as the measure of a parameter showing the least deviation when subject to different kinds of variation. The variation can here be either external, i.e. something mechanical affecting the sounds produced when speaking, such as noise created by:

- recording equipment
- duration (stability)
- noisy environment

or internal, such as:

- emotional state
- speech liveliness variation
- speaking style
- disguise
    - dialect/sociolect
    - voice quality.

From an engineer's perspective, noise often refers only to environmental noise, such as in poor quality recordings (Kimura, Ashida & Niyada, 2004; Nakasone & Beck, 2001; Reynolds et al., 2000; Reynolds, 2003) or mismatched recordings, i.e. recordings made on different recording or transmission media (Alexander, Dessimoz, Botti & Drygajlo, 2005). All of this is what is referred to as noise here. However, noise that affects parameter values conveying variation in phonetic and linguistic studies can also be given a wider definition. Between-speaker variation in perceptual studies, for example, may be seen as noise that affects robustness (Clopper & Pisoni, 2004), as can within-speaker variation due to changes in the speaker's emotional state (Doherty & Hollien, 1978), speech liveliness (Traunmüller & Eriksson, 1995) or level of vocal effort (Jessen, Köster, & Gfroerer, 2005). This wider definition is used here.

There are patents covering robust speaker recognition or robust pattern recognition, where robustness is not explicitly defined (Pilz, 2006). In forensic speaker comparison, the concept of robustness is used in conjunction with methodology (Drygajlo et al., 2003; Gomez, Alvarez, Mazaira, Fernandez & Rodellar, 2007). Robust statistics have several different definitions depending on whether one applies a frequentist or Bayesian approach, i.e. depending on the choice of probabilistic or non-probabilistic methods (Huber, 2011).

The concept of robustness is often used in speaker-comparison contexts when referring to the discriminative power of a parameter (Gomez et al., 2007; Lindsey & Hirson, 1999). In this thesis, it is important to differentiate between robustness of the measures involved in measuring a parameter vs. assessing its discriminability, i.e. the within- vs. between-speaker variation.

The investigation of the amount of statistical data needed to assess the successfulness of a parameter as a discriminator between speakers has not been sufficiently covered. There are many suggestions and theoretical claims, but often these are not tested in a proper forensic context, such as noisy and/or band limited environment recordings and the comparison between mismatched conditions. The term “mismatched conditions” here primarily refers to the comparison between recordings made on different recording equipment and/or in different environments but is also used to refer to recordings made in two totally different speaking modalities. A clear example containing both these usages of the term is the comparison between one short recording from a bank environment where an unknown speaker shouts commands to a cashier, and one high-quality recording made in a studio environment of an interview in a low-voiced modality. It is difficult to decide what the proper data requirements to make a robust forensic speaker comparison are (Schwartz et al., 2011b). First of all, a question at issue from an investigating officer’s perspective is how long a recording of an unknown voice should be. This question does not, however, have an easy answer, and it cannot be given in seconds. It depends on what is on the recording. The question generates several follow-up questions such as: What is the audio quality of the recording? What kind of speaker is talking and how (i.e. what modality, speaking style etc.)? Are there any peculiarities in the unknown speech or voice? This has to be investigated first. For phonetic features, it is sometimes possible to derive a lot of information from a short recording and to some extent by the use of an automatic system. It is, however, unlikely that the conclusions drawn are very reliable.

5.3.3 Robustness of forensic phonetic parameters

An important issue when discussing robustness in a forensic setting is the confusion about the use of the concept itself, i.e. that there is a difference between robustness and the usefulness of a parameter. A parameter can have great precision and performance for comparing speakers, but that does not necessarily imply that the parameter is robust against different kinds of variation. Nolan (1983) attempted to define criteria for suitable parameters to be used in forensic speaker comparison. A robust and useful parameter should adhere to these criteria:

1. Availability.
2. Measurability.
3. Robustness in transmission
4. Resistance to attempted disguise or mimicry.
5. High between-speaker variability.
6. Low within-speaker variability.

Availability and measurability are of course prerequisites, i.e. parameters for analysis must be available and measurable. Criteria 3 and 4 capture the issues regarding robustness to noise, both from internal and external sources (as discussed in section 5.3.2).

Criteria 5 and 6 are in a sense the results of an analysis, but can be predicted in advance from prior knowledge. That is, previous research and casework experience yield an understanding of which parameters serve to discriminate speakers well in a given community. The power of discrimination can be measured using for example the F-ratio. The ratio measures whether the variances are equal or become higher if the between-speaker variability is greater than the within-speaker variability, i.e. the higher ratio the better (Kinoshita, 2001; Kirkland, 2003; Alderman, 2005).

To establish which measure for a parameter, such as F0, is the most robust, one can create tests in different ways. By adding different kinds of noise (i.e. variation) to the parameter and then using different measures, it is possible to investigate which one produces least deviation from a gold standard, i.e. generates the least within-speaker variation. The gold standard could be the true value if that is known, or the value without any noise-producing variation. When the most robust measure is found, it is natural to proceed to test the variability of the measure from a forensic perspective, meaning by calculating the within and between variation for a suitably large reference population. For a forensic analysis, one can attempt to calculate the differences between samples and a relevant reference population. It is important to always consider the differences and similarities between the samples, but by using the reference population typicality is considered too, i.e. how common that specific similarity or difference is compared to a similar pool of speakers. A suitable framework for forensic evaluation is the Bayesian likelihood ratio framework. If it is possible to record or collect enough suitable reference data, it is possible to calculate likelihood distributions for a parameter. However, sufficient forensically realistic data is hard to find and tedious work to collect. A likelihood ratio (LR) between two competing hypotheses, often referred to as the prosecution and the defence hypothesis (Champod & Evett, 2000; Champod & Meuwly, 2000; Kinoshita, 2005), can then be calculated using the data from known, disputed and reference samples and be presented as the result of a forensic analysis. The work of estimating the discriminative power of speech parameters such as formants and F0 have been attempted by several researchers. To be able to correctly express an outcome from a test using formants, some way of calculating the LR in a multivariate way is needed (Rose, 2010). It is also possible to use discriminant analysis to investigate the power of a parameter using a general LR as a discriminant function (Evett, Scranage & Pinchin, 1993; Kinoshita, 2001; Meuwly & Drygajlo, 2001; McDougall, 2004, 2006). However, it is not clear how to calculate the LR, even though there are suggestions (Aitken & Lucy, 2004; Morrison, 2009). In Rose (2002) an attempt is made using a univariate formula developed for glass fragment comparisons (Aitken, 1995) applied to analysis of fundamental frequency (F0). Kinoshita (2005) showed that Lindley's formula, which leaves out the multi-session variation present in speech data, works, as it produces realistic LRs for F0. The LRs produced were close to unity and the conclusion was that F0 will not produce evidence that strongly supports any hypothesis within forensic speaker comparison. Measures used for fundamental frequency such as mean and standard deviation have been shown to be very unstable (Traunmüller & Eriksson, 1995) and using a univariate model is a very crude way to model F0 variation, which is the main weakness of the formula. When the strength of another acoustic measurement such as vowel formants is measured, the aspect of variation becomes even more complicated and new ways were tested to calculate a LR. Alderman (2005) attempts to apply one of Lindley's formulae to formant data, with some degree of success. To correctly cover the type of variation, further elaboration on how to calculate LRs for formants were tested using Multivariate Kernel Density (MVKD) published in Aitken & Lucy (2004). This is another formula intended for the forensic analysis of glass fragments. It has been tested for forensic voice comparison purposes in calculating multivariate LRs in several studies (Ishihara & Kinoshita, 2008; Rose, Kinoshita & Ishihara, 2008; Kinoshita, Ishihara & Rose, 2009; Morrison, 2009; Morrison & Kinoshita, 2008; Rose, 2006) and was recently compared to a system using a UBM-GMM approach from automatic voice recognition (Reynolds, 1995, cited in Morrison, 2011). The different aspects of expressing evidence and the application of the formulae will be discussed to some extent in the following sections.

5.3.2.1 Fundamental Frequency

Fundamental frequency has been a central physical feature in the area of forensic speaker comparison for a long time. As descriptors of individual differences in fundamental frequency, long-term distribution measures such as arithmetical mean and standard deviation have often been suggested (Rose, 2002). These measures depend on duration, however, and there is no general agreement on what minimum duration is required to yield reliable results. (Horii, 1975) suggests that recordings should exceed 14 seconds. In other studies, ranges from 60 seconds (Nolan, 1983) up to 2 minutes (Baldwin & French, 1990) are suggested as a minimum. Rose (2002) reports that F0 measurements "/.../ for seven Chinese dialect speakers stabilised very much earlier than after 60 seconds", the duration suggested by Nolan (1983), implying that the values may be language specific. Braun (1995) further discusses the problem of minimum duration, suggesting that it is dependent upon psychological or physiological factors, but mentions that 15-20 seconds is sufficient "/.../ if the communicative behaviour may be considered 'normal'". Positive skewing of the F0 distribution for a speaker is typical (Jassem et al., 1973). This means that the values will not be symmetrically distributed around the mean, but more values will end up on the higher end than the lower. Since it has been shown that the standard deviation increases with differences in speech liveliness (Traunmüller & Eriksson, 1995), this factor will influence the mean since the distribution is shifted upwards with more liveliness and thereby automatically increases the mean. In order to make the measure insensitive to this kind of variation, a different measure needs to be used. Braun (1995) mentions noise (in a wider sense) as something affecting the measurement of F0. In forensic cases, the questioned recordings and the reference recording are often different both with respect to speech style and audio quality. Using the traditional measures for describing fundamental frequency level, such as mean or median values, may therefore yield misleading results. The intra-speaker variation is affected by paralinguistic and other factors as has been very well described. Braun (1995) categorises the variations as technical, physiological or psychological factors. Tape speed, which perhaps surprisingly is still an issue for forensic samples, and sample size are examples of technical factors; smoking and age are examples of physiological factors, while emotional state and background noise are examples of psychological factors. In spite of all these sources of variation, fundamental frequency has nevertheless been studied a lot and claims have been made that it can be a successful forensic phonetic parameter (Atal, 1972; Baldwin & French, 1990; Hollien, 1990; Künzel, 1987; Nolan, 1983; Rose, 2002). Another obvious problem is that F0 (mean) has a roughly normal distribution across the population (Lindh, 2006), hence its forensic value is inherently limited and could only offer any contribution FSC when extreme values are present. Therefore other properties of F0 are examined here and not simply the mean.

5.3.2.2 Formants

One of the main types of analysis performed by examiners within forensic speaker comparison has traditionally been vowel formant analysis. A formant can be described as the centre peak of the acoustic energy generated by the resonant frequencies resulting from how a speaker's supralaryngeal cavities are configured. The three lowest frequency formants are the ones mostly connected to how we articulate and perceive vowels. The formants are not just connected to the way our cavities are configured, but also connected to our biological build, i.e. formants are connected to the structure (anatomy and physiology) of the supralaryngeal articulators.

Since these physically connected frequencies with the highest amplitude are rather variable, a range of sources of intra-speaker variation have to be dealt with when measuring them. Measuring and comparing formant frequencies is problematic. The research on formants in forensic phonetics and the effect of measuring them started earlier than the research and contributions of Nolan (1983). LaRiviere (1975) showed that formants and F0 contributed to aural speaker comparison judgments using a small data set of vowels. Further, in a study of American English, Goldstein (1976) calculated F-ratios using formants to see which parameter was the best discriminator, and F2 in [r] was shown to be the most effective feature. An identification task using only the two best features showed 12 errors out of 80 tests. The tests by Murry and Singh (1980) showed that listeners use vowel information mainly to discriminate between vowel pronunciation by different speakers, while the results of Brown (1981) indicate that voice similarity (on synthetic speech) was judged on the basis of F0, formant mean and bandwidth. Other intra-speaker effects influence formant values and trajectories, such as speaking rate (Imaizumi & Kiritani, 1989), timing (Eriksson & Wretling, 1997), context variation (Ingram et al., 1996; Mullennix, Johnson & Topcu, 1991) and vocal effort (Traunmüller & Eriksson, 2000). Nolan and Oh (1996) made an attempt to spot differences between twins using formants, but without much success. A few experiments have been made to correlate formants (and other parameters) with speaker size. Only weak correlations were obtained in respect of height (Greisbach, 1999). Little or no correlation was found by Dommelen & Moxness (1995).

The turning point in using formant measurements in speaker-comparison casework came in the nineties, when acoustic analysis became more important, and crucially much easier due to technical development, in conjunction with auditory analysis (Ellis, 1990; French, 1994; Nolan, 1990). Software development speeded up the process and in the UK, measuring formants was for a time (ca. 2007-12) in effect obligatory when performing speaker comparison[9]. Quite early on, it became obvious that comparing recordings from taped interviews and intercepted telephone conversations was problematic (Rathborn, Bull & Clifford, 1981). More recently, GSM telephony has become another problem (Harrison, 2004), but in the beginning no difference in

[9] The Queen v Anthony O'Doherty 19/4/02 ref: NICB3173. Court of Criminal Appeal Northern Ireland.

performance for speaker verification systems was reported (Kuitert & Boves, 1997). It was also shown that listeners performed better than automatic systems (Schmidt-Nielsen & Crystal, 1998). The effects of the telephone filter were tested as a consequence, and Künzel (2001) found that the lower cut-off frequency for the bandpass filter creates an artificial upward shift for the centre frequencies. The effect was shown for all German vowels tested except /a/, for which the higher F1 was unaffected. The effects were then further discussed in a forensic (and somewhat dialectal) perspective. The study was interpreted as an attempt to exclude the measurements of formants for forensic speaker comparison purposes, but that was however not the purpose (Nolan, 2002b). The Künzel (2001) study only shows a significant effect on F1, which implies that caution should be exercised when measuring that formant, while higher formants would presumably remain unaffected (Künzel, 2002). Byrne and Foulkes (2004) measured the mobile phone effect on formants and the results showed similar effects on F1, less or no significant overall effect on F2 (even though most tokens were affected in some way), while F3 could be affected if high enough to be affected by the upper cut-off filter effects. The importance of measuring formants despite the effects is discussed in the frame of a forensic case in Nolan & Grigoras (2005). Further elaborated tests on GSM transmitted speech was tested in Guillemin & Watson (2008) and significant effects were found, especially when formant frequencies were automatically extracted. It is still not clear how robust the extraction of formant measurements is or which measurements should be extracted for comparison between vowels produced by different speakers, i.e. two criteria are not fulfilled for a reliable parameter in forensic casework.

The solution in casework for the variation of the parameter has so far been to use manual intervention when measuring formant frequencies. How reliable it is to measure formants manually is not clear from the literature. However, recent research (Harrison, 2013) has shown that by using an optimal set of LPC (Linear Predictive Coding) coefficients for the specific vowel type, by being aware of possible errors and by using manual correction, errors can be kept to a minimum. In Ishihara & Kinoshita (2008); Rose et al. (2008); Kinoshita et al. (2009); Morrison (2009); Morrison & Kinoshita (2008) and Rose (2006), formants are measured with manual checking and the discriminative value of the parameter is measured by strength of possible evidence using multivariate likelihood ratios. A multivariate likelihood ratio is however an improvement over the univariate likelihood ratio first used for formants in Kinoshita (2001).

Formants were early on considered to be one of the most promising parameters for forensic speaker comparison and they were also shown to have discriminative power using the F-ratio (Wolf, 1972), but unfortunately only for very small data sets. Kinoshita (2001) also used this approach (i.e. F-ratio) to select the most discriminative values for different vowels produced by 13 male speakers of standard Japanese on two occasions. The discriminative power of the measure is, however, difficult to calculate using such a small data set. The test of formants as a discriminative parameter is further elaborated in Alderman (2005) using the Bernard data for Australian vowels as a background model of the formant distribution. However, that data from 170 male speakers was manually measured in the late sixties. The error rates for the measurements are unknown (Bernard, 1970). Alderman (2005) further describes the lack of intra-variation data as a setback of the background data (2-3 tokens per vowel per speaker) as well as the non-contemporaneity. In a forensic perspective it must be concluded that it is unclear how well formant measurements can discriminate speakers. The estimates presented in the different studies show LRs which are rather low, but of course could be a helpful ingredient in combination with other tools performing speaker comparison. Instead of using single vowel formants several studies have been performed using Long-Term-Formant (LTF) analysis and applied in a similar manner as automatic systems (Gold, French & Harrison, 2013; Jessen et al., 2013). This means that even though a large effort has been made with regard to research on the use of formant frequencies in casework, there is still a lot to be done to make them a useful parameter. Even the combination of this kind of parameter with an automatic system and other analyses produces very little gain in performance for speaker comparison considering the heavy workload involved in segmenting and estimating (unless automatic extraction is used).

5.3.2.3 Other Phonetic and Acoustic Features

Over the years other features have been investigated for the purpose of discriminating speakers. Rhythm, intonation patterns, rate of speech and pausing are some examples. Rhythm and pitch were tested on listeners in Dommelen (1987); the results showed that the listeners did use the parameters, but that it depended on how atypical the specific speaker was. The experiment was further elaborated in Dommelen (1990), where pitch was found to be very relevant in a speaker discrimination task and speech rhythm least significant. Rhythm has also been investigated to some extent for laughter (Bachorowski & Owren, 2001; Bachorowski & Smrkovski, 2001; Kipper & Todt, 2001, 2003a, 2003b; Mowrer, LaPointe & Case, 1987; Owren & Bachorowski, 2003; Poyatos, 1993; Provine, 1993; Smrkovski & Bachorowski, 2002), research that shows stereotypicality for speakers (Provine & Yong, 1991) and which therefore concludes that rhythm can have potential as a discriminatory feature. The pairwise variability index (PVI) for measuring speech rhythm or syllable timing of speech (Grabe & Low, 2002) should also be mentioned. More recently a group in Switzerland are investigating several different measurements of speech rhythm (Dellwo, Leemann & Kolly, 2015; Leemann, Kolly & Dellwo, 2014).

When it comes to a feature like speaking tempo, there are several relevant studies, not always concerned with forensic phonetics. In Shipp et al. (1992), temporal features and their correlation to perceived age are investigated. Johnson, Ladefoged & Lindau (1993) looked at the specific articulatory patterns in the production of vowels, while Künzel et al. (1995) investigated the relation between tempo, loudness and F0, specifically in the forensic context. Künzel (1997) then tried to pin down a more practical approach, measuring pauses (using a minimum of 100ms) and speaking/articulation rate in three different conditions (spontaneous interview, spontaneous conversation and reading neutral text). Speaking rate (SR) can be defined as the number of words or syllables produced per unit of time of choice (depending on your data). Articulation rate (AR) is the number of syllables per unit of speaking time (i.e. the recorded speech minus pauses).

Other interesting parameters investigated are features such as jitter for quantifying voice quality or level of hoarseness (Wagner, 1995). Unless very sophisticated automatic methods for extraction (such as a high performing phone recogniser) are used, these parameters are time consuming (PVI) or difficult (jitter) to measure in noisy environments. A few important studies in the field of phonology (Moosmüller, 1997) and intonation (Nolan, 2002a) have also been conducted.

5.4 Tools and databases

Automatic methods are increasingly being used in forensic phonetic casework, but often in combination with perceptual acoustic methods. The NIST speaker recognition evaluation campaign started as far back as 1996 with the purpose of driving the technology of text-independent speaker recognition forward, but, also aimed to test the performance of the state-of-the-art approach and to discover the most promising algorithms and new technological advances[10]. The aim of the test is to have an evaluation at least every second year, and some tools are provided by NIST to facilitate the presentation of the results and handling of the data (Martin & Przybocki, 1999). A few labs have been evaluating their developments since the very start with improving performance results over the years. The labs that have evaluated their system for a longer period of time have generally performed best in the evaluation. To develop a working system three basic parts are needed:

1. Parameter extraction - feature extraction from audio (most often MFCCs, Mel-Frequency Cepstral Coefficients, i.e. the coefficients of an MFC vector).
2. Training - to create models from the extracted parameters.
3. Testing calculation of distances between parameters extracted and models created.

There are many different ways to proceed through these different steps, which are the steps used for the system applied in the different experiments involved in the work presented here. It is possible to download and use all programs described and presented here as they are distributed as open source.

5.4.1 The Terminology for automatic systems

It is difficult to grasp the terminology in this area of research because terms are used differently depending on whom you are talking to or what you are reading. It depends on the person’s background and the attitude toward certain methods. In this thesis, there are several terms that need a detailed description or definition, especially when it comes to automatic systems. These are given in this section. In the commercial sector, automatic speaker recognition has mainly been divided into two different branches usually described as *text dependent*, a system which is dependent on what is being uttered by a user when deciding acceptance or rejection, or text *independent,* where a system does not depend on what is being uttered. The term "text dependent" means that the system is expecting a specific set of words to be uttered by a preregistered user and these words combined with the parameters of the voice will be accepted or rejected. A "text independent" system is not expecting any particular set of words, but will analyse and compare anything uttered and compare it to stored speaker models. Both these types of systems presume that someone is claiming an identity to be accepted by the system, which is normally categorised as *speaker verification,* an AVC or ASC system with an

[10] http://www.nist.gov/speech/tests/sre/ fetched Jan 12, 2009.

incorporated threshold to verify, i.e. accept or reject a user. Speaker verification is a term normally used for commercial systems such as gatekeepers or phone access systems etc. Such systems differ from others in the sense that they have a certain threshold for acceptance vs. rejection depending on a *likelihood ratio* (LR) score, i.e. the ratio between two competing hypotheses as a part of a Bayesian probability test, that is, a score that has to be chosen or calculated on a test data set in advance. The threshold is connected to the security level of the system, adjusted in line with whether it is more important to avoid false rejections or false acceptance.

The term *automatic speaker identification* is most often used to describe a system used for attempts to identify unknown speakers, which is the case in forensic casework. However, due to the probabilistic nature of forensic research, the term identification is inappropriate, which is why *automatic voice comparison* (AVC) is introduced and used here instead. AVC refers to a system that ignores linguistic and/or phonetic information and solely compares the acoustic vectors in hand by statistical measures. It is an estimated biometric calculation of what is in the investigated signal. The enormous variation voice is affected by in different situations from the same or different voices means an automatic analysis should be completed with several other phonetic and linguistic features. When using the term *automatic speaker comparison,* the system should refer to a system that models some sort of phonetic, phonological and linguistic features.

When comparing voices in a system, there are situations called open or closed sets. An *open set* is where a whole language community constitutes the possible options for being an unknown voice, while a *closed set* implies that a known and finite set of possible voices. A calculation of which one is the closest to the unknown voice is then necessary.

5.4.2 Parameterization

The parameterization of early attempts at automatic speaker verification meant classification of well-known parameters such as fundamental frequency (Atal, 1972) or linear prediction coefficients (Atal, 1974). The very first systems were otherwise averaged output of analogue filters and visual matching of short spectrograms (Pruzansky, 1963), sometimes involving humans (the voiceprint approach) (Bricker & Pruzansky, 1966). A prototype system was built in 1976 by Texas Instruments and tested by the U.S. Air Force (Haberman & Fejfar, 1976). In the same year, Atal made a summary of the techniques and parameterization tested so far (Atal, 1976). The group of speech technology researchers that started the evaluations at NIST drove the development forward and (as mentioned) in 1996 the annual evaluations of systems started[11]. The approach has diachronically been to try and model the voice in different manners and with the tools available at the time. Early research showed that the intra-variability for a parameter such as F0 is somewhat discriminative for a small set of speakers, but that a useful model of the vocal tract would most likely be more attractive (Atal, 1972). The formants then became obvious candidates for being valuable to model. When automatic systems are discussed, one must generally be aware that they are merely statistical machines and the result will always depend on what parameter one feeds them with.

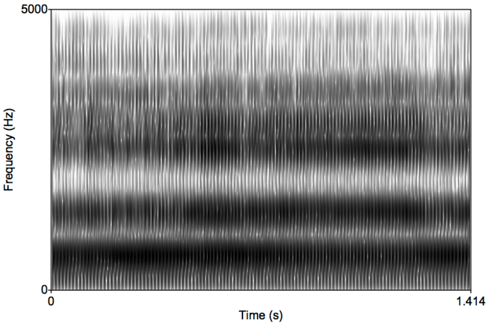
Extraction of features in speech contains several layers of problems, starting with the audio quality of (most often) digitalised audio. In forensic audio, one will run into many different kinds of audio recordings made in many different environments. The biggest problems come with the different compressed formats such as mp3. To compress audio and decrease file size, information is removed. More or less successful algorithms succeed in removing redundant data, most often defined as information not detectable by the ear ("beyond the auditory resolution ability")[12]. However, some of the measurable features in speech might also be removed, such as formant frequency information. To be able to handle the different formats, the systems must have technical ways of normalising or taking care of only the relevant (i.e. information rich) parts of the speech data, not necessarily the same information used when interpreting the auditory signal. A second, just as relevant, problem is the variation induced by different microphones, which has to be modelled. After considering the variation from the different audio formats and recording equipment, there is the environmental "noise" from where the recording took place, which has to be taken into account.

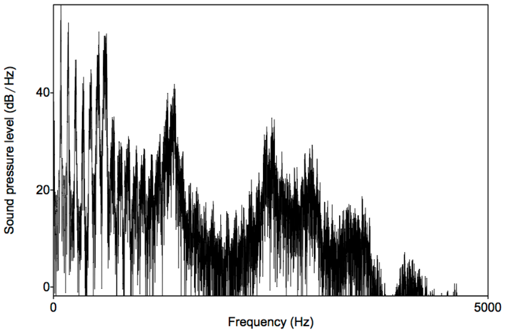
The most common features in automatic voice comparison systems are as mentioned Mel-frequency cepstral coefficients (MFCC). These coefficients create a Mel-frequency cepstrum (MFC), which is a representation of the power spectrum of a sound (slice) by a transformation of the log power spectrum on a Mel-frequency scale[13] of a signal. This cepstrum is a short-term power spectrum of a signal (sometimes referred to as a non-linear spectrum of a spectrum) originating from a variation of Fourier transforms (FT), a mathematical operation that

[11] http://www.itl.nist.gov/iad/mig/tests/sre/
[12] http://en.wikipedia.org/wiki/MP3
[13] http://en.wikipedia.org/wiki/Mel-frequency_cepstrum

decomposes a signal into its constituent frequencies. $\mathrm{FT}_{\mathrm{log}}\ \mathrm{FT}\tau$ (Bogert, Healy & Tukey, 1963), where τ is a frame window of audio (Oppenheim & Schafer, 2004) (most often 10ms long). The nonlinearity is derived from the Mel scale,[14] which is a perceptual scale based on listener judgments of equal distances between tones (Stevens, Volkmann & Newman, 1937). The coefficients therefore work in what the originators would call the quefrency domain (Bogert et al., 1963) and formants are related to lower coefficients and the fundamental to the higher ones (Gannert, 2007). There are other features such as the Perceptual Linear Predictive speech analysis (PLP) (Hermansky, 1990), which is based on a similar technique but is related to the perceptual Bark scale, which is based on 24 corresponding critical bands of hearing[15] (Zwicker, 1961). In evaluations, PLPs have been shown to perform similarly to MFCCs (Psutka, Müller & Psutka, 2001).

Classical visualisation of speech is made using a spectrogram,[16] which is a time-varying representation of spectra. Time is on the x-axis, frequency on the y-axis and the energy is displayed in a grey scale, i.e. the darker the more energy. The spectrum is the digital Fourier transform of a signal.[17] To get a better overview of the spectral shape of a sound over a time period, it is possible to average the energy in every frequency band with a certain bandwidth: this is called a Long Term Average Spectrum (LTAS).[18] In comparison with the spectrogram, the MFC can be displayed in the same manner by plotting the grey scale of the level of each coefficient for each frame on the time scale.





[14] http://en.wikipedia.org/wiki/Mel_scale
[15] http://en.wikipedia.org/wiki/Bark_scale
[16] http://en.wikipedia.org/wiki/Spectrogram
[17] http://www.phys.unsw.edu.au/jw/sound.spectrum.html
[18] http://www.fon.hum.uva.nl/praat/manual/Ltas.html

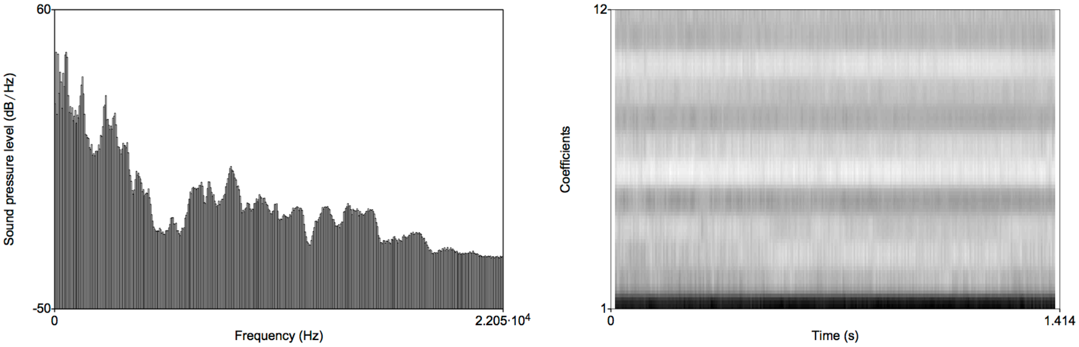


Figure 2. Spectrogram, spectrum, Long term average spectrum (50Hz bins) and 12 coefficient MFCCs of a sustained schwa, neutral vowel, recorded by the author.

Before actual extraction of features is performed in a generic automatic system, an energy detection is performed. Usually this means finding the speech data in a signal and ignoring the silent parts. In the parameterization step, a normalisation procedure to avoid the influence of environmental noise is performed using *cepstral subtraction*. Feature normalization can be used to reduce the mismatch between signals recorded in different conditions. Normalization consists of mean removal and eventually variance normalization[19] (Guillaume, 2004). This is a rather crude normalisation as it presumes that environmental characteristics over a signal are constant.

5.4.2.1 Modelling voices

When the extraction of the basic parameters has been performed, a system has to model single voices as well as groups of voices to be able to measure distances for both similarity and typicality. The most generic way to do this is to use the UBM-GMM technique (Reynolds et al., 2000). A *Gaussian Mixture Model* (GMM) is a kind of clustering technique where a mixture model[20] is a probabilistic unsupervised technique to represent the presence of subpopulations within an overall population. This is done using *expectation maximisation* (EM).[21] In statistics, an expectation maximization algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. In statistics, latent variables (as opposed to observable variables), are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (i.e. directly measured)[22]

[19] http://www.irisa.fr/metiss/guig/spro/spro-4.0.1/spro_2.html
[20] http://en.wikipedia.org/wiki/Mixture_model
[21] http://en.wikipedia.org/wiki/Expectation-maximization_algorithm
[22] http://en.wikipedia.org/wiki/Latent_variable

(here the mixture models). The *universal background model* (UBM) is a large GMM trained to represent the voice-independent distribution of features in an AVC system. This large GMM is trained with an expectation-maximization algorithm (EM) (ten iterations) to represent the speaker-independent distribution of features. The idea is to cover the general characteristics of a given population (i.e. the voices you have chosen to represent your system's world of features). This large model is used as a starting point for training voice models tested by the system. To train these voice models (not present in the UBM), a different technique is applied called *maximum a posteriori.*[23] In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is a mode of the posterior distribution. A more detailed overview of the technique can be found in Gannert (2007).

5.4.3 General Performance and Evaluation of Automatic Systems

There are different ways to evaluate an automatic system, especially when it comes to the text-independent systems, which are generally used in a forensic context. In this context, it must be considered how well the systems perform under different conditions to be useful. Doddington (1985) highlights this question as early as the mid-eighties. He further elaborates on the difficulties of mismatched conditions and the NIST evaluation 1998 (Doddington, Przybocki, Martin & Reynolds, 2001) using DET (Detection Error Trade-off). A detection error trade-off graph is a graphical plot of error rates for binary classification systems, plotting false reject rate vs. false accept rate curves[24], which plot missed detections and false alarms for a system's performance. The DET curve is one common way to compare different systems and is the follow-up to the ROC curve[25] (receiver operating characteristic). A receiver operating characteristic (ROC) curve is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate for a binary classifier system as its discrimination threshold is varied. It is used to plot false alarm probability on the X-axis and detection rate on the Y-axis (linear scale) (Fawcett, 2004).

[23] http://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation

[24] http://en.wikipedia.org/wiki/Detection_Error_Tradeoff

[25] http://en.wikipedia.org/wiki/Receiver_operating_characteristic

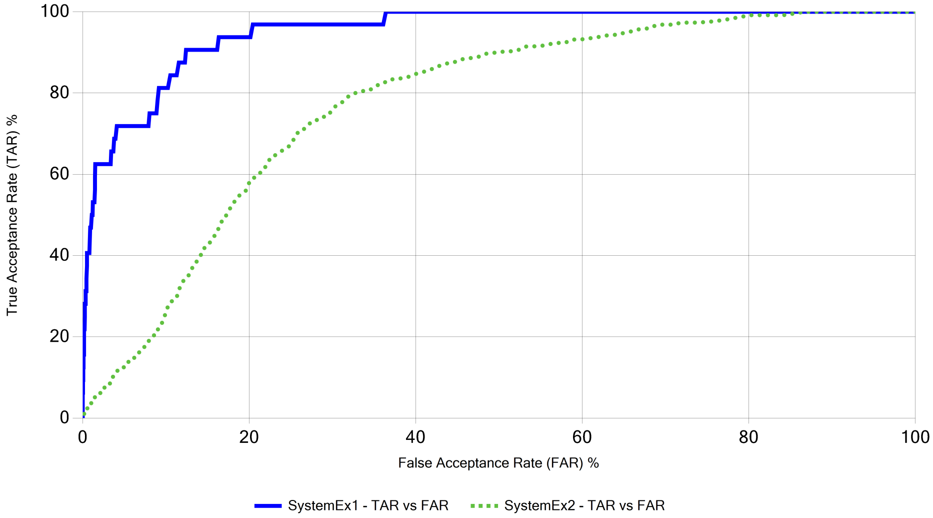


Figure 3. Example of ROC curve comparison between two example systems.

Martin, Doddington, Kamm, Ordowski & Przybocki (1997) argue that the DET curve gives better visual information on the normality of the scores (normality equals straighter line), which gives information on the error type involved. This is the normal way to report improvement or testing in speaker-verification research. The point where the false alarm probability and the miss probability are the same is called the Equal Error Rate (EER)[26], i.e. the rate at which both accept and reject errors are equal. When one wants to evaluate a system and has access to a database, there are packages provided by NIST to plot the results and compare different versions of your system (or compare against other systems) in Matlab[27] or Gnuplot[28] as well as a lot of other useful tools.[29] In this way, system performances are compared both within groups and between groups during the NIST evaluations. Some other evaluations have been performed such as EVALITA[30] and there has been evaluation with a more forensic approach by a Dutch initiative (van Leeuwen & Bouten, 2004). The best performing system in that forensic evaluation showed an EER of 12.1% (where the test material was wiretapped recordings from real Dutch police investigations) and the best result was obtained from the test using 15-second test recordings and 60 seconds of training data. Since then, systems have improved. More difficult tasks have been tested. Mismatched conditions have been tested too, which, as mentioned, are

[26] http://www.griaulebiometrics.com/en-us/book/understanding-biometrics/evaluation/accuracy/matching/interest/equal
[27] http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm
[28] http://www.itl.nist.gov/iad/mig/tools/gnu_detwaretarZ.htm
[29] http://www.itl.nist.gov/iad/mig/tools/
[30] http://evalita.fbk.eu/

common in forensic material. Different audio formats, some compressed and different channels such as landline telephone recordings vs. interrogation room recordings and GSM recordings have been tested. From a technical point a view, several sessions of recordings from the same speaker in different mismatched conditions would be needed to appropriately model one single speaker. However, different compensation techniques have shown promising results both theoretically, in publications such as Alexander (2005) and practically implemented in the open-source system used for the studies in this thesis, in publications such as Matrouf, Scheffer, Fauve and Bonastre (2007). In a follow-up, it was even shown that some of the mismatching could be compensated by using a weight-based factor analysis to model inter-session variability, regardless of whether you have same speaker data. With different normalisation techniques an EER as low as 1.62% was achieved using the NIST 2005 speaker-recognition evaluation material (McLaren, Matrouf, Vogt & Bonastre, 2008).

When it comes to forensic systems, it is somewhat different, however. First of all a measure of "goodness" and calibration is needed. The C_{llr} (Log Likelihood Ratio Cost) is such a measure (Brümmer & du Preez, 2006). C_{llr} is a measure of the overall strength of support of all the log likelihood ratios for all comparisons in an evaluation of a system and penalises errors according to their magnitude. The lower the C_{llr} the better validated the system is. Usually Tippett plots then graphically show how well a system separates the comparisons. An example is given in Figure 4. Same speaker comparisons (SS) are represented by the blue curve (to the right) and the different speaker comparisons (DS) by the red (to the left). The x-axis represents the log LR scores and the y-axis the cumulative proportion of comparisons at or beyond the x-axis value. In the figure the SS values to the left of the 0 point and the DS values to the right of the 0 point are discrimination errors. The EER can be inferred by the crossover point, i.e. in Figure 4 the EER is approximately 10%.

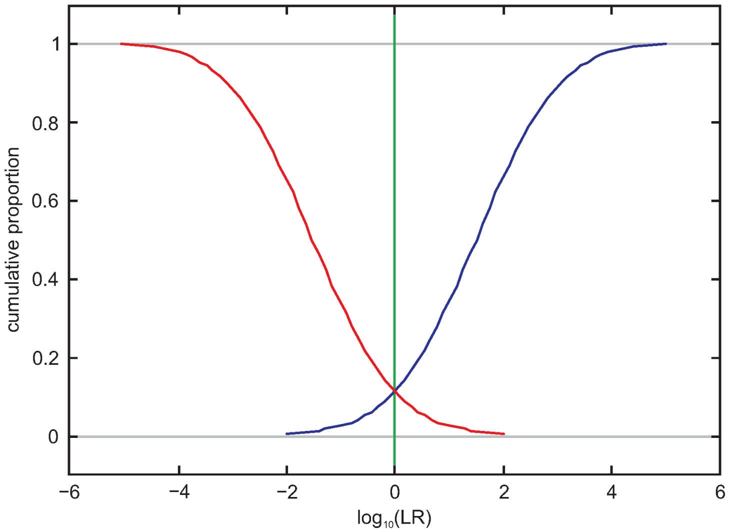


Figure 4. Example of a Tippett Plot.

To be able to demonstrate proper reliability and validity of a forensic system this kind of evaluation is suggested (Morrison, 2011). The evaluations provide us with important information about the improvements and drawbacks of automatic systems. Commercial systems can use the evaluations as a quality check and marketing material if their performance is good. However, this type of evaluation does not establish what types of mistakes are made, whether the same mistakes persist across different systems or after improving a single system. Many engineering studies focus on the fact that different statistical techniques are applied and performance is affected, but many times without a prior theoretically based hypothesis. Most improvements are also hard to compare since a different system might have been used to implement the same technique but get a different result, possibly depending on system architecture or the test material. One way to deal with that kind of problem is evaluating new techniques and algorithms using open-source software. The following subsection presents such a system, which is also the software used for the studies in this thesis.

5.4.4 ALIZE - an open-source platform for biometrics

As one way of improving system development and evaluation the Avignon group made available their system, ALIZE, under an open-source LGPL licence[31] (Bonastre et al., 2005; 2008). ALIZE

[31] http://www.gnu.org/licenses/lgpl.html

is a platform for building biometric applications, demonstrators and/or research. The platform was originally based on a generic UBM-GMM approach, but also contains functionality for using very recent advances in speaker recognition, such as latent factor analysis, SVM supervectors, unsupervised adaptation and more recently the i-vector technique. While the performance of the basic technology is tested within the NIST evaluation, several other applications and tools have been developed. ALIZE is the umbrella for all developed packages that could be included in the users own application or framework. A thorough description of an experimental setup is given in (Mayoue, 2008). The system used in all experiments for this thesis has a similar setup.

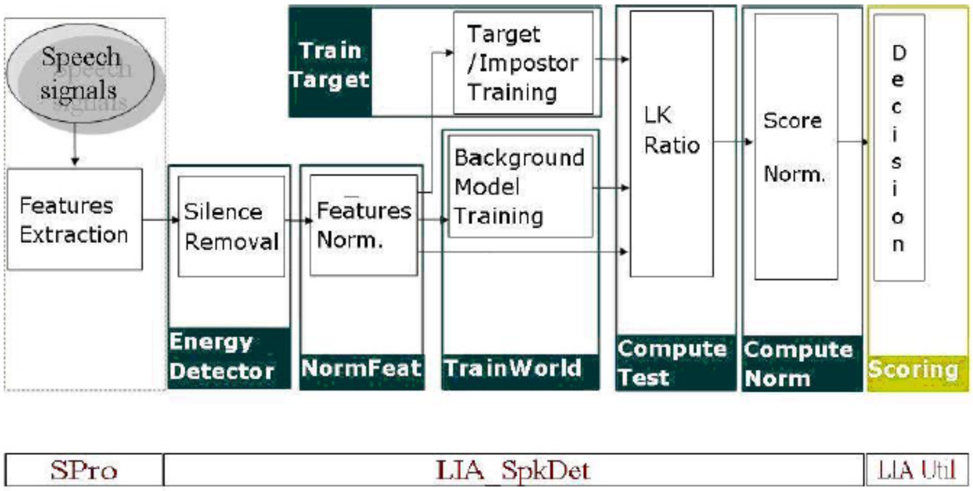


Figure 5. Standard setup for an experiment using ALIZE (reproduced from Mayoue, 2007).

The feature extraction can effectively be handled by SPro, an open-source signal processing toolkit (Guillaume, 2004). In this case, MFCCs are used, calculated on 20ms Hamming windows with a 10ms shift. For every frame a 19-element vector is calculated together with first and second (acceleration) order deltas. In addition energy and delta energy are used. This is applied to wav pcm files with a sampling rate of 16 kHz without bandwidth limitation. This gives a 19 (MFCCs) x 3 (with 1st and 2nd order derivatives) +2 (energy + delta energy) = size 59 feature vectors. In ALIZE, energy detection is applied using X of the windows with most energy. This procedure is, however, not used for our system. Instead a simple energy detection is applied in a pre-processing step using all frames meeting criteria >-25dB and >100ms. This was applied using a script implemented in Praat (Boersma & Weenink, 2007). The retained frames' features are then normalised on both mean and variance. The Gaussian Mixture (also so-called “world”)

Model (UBM) is trained in beforehand using the EM algorithm and 512 mixtures[32] on 2 minutes net of spontaneous speech from each of 628 male speakers in the SweDia dialect database (Lindh & Eriksson, 2009). Training the target models was done by adapting the parameters of the UBM using the speech from the target and maximum a posteriori (MAP) estimation. For the target model training, again 512 mixtures are used and log energy coefficient discarded. The MAP method used is the MAPOccDep approach (the random variable to be estimated is computed by a linear combination of its value in the world model and the value obtained by an EM algorithm on the data). This method takes into account the a posteriori probability n for each Gaussian. The weights of this combination are provided by the option MAPRegFactor r (n/n+r for the world model and 1-n/n+r for the target model) (Mayoue, 2007).

5.4.4.1 Integration of Phonetics and Linguistics in AVC - A Forensic Perspective

A major difficulty for closer integration of phonetic and ASR methods lies in the fact that one must know which parameters are dependent or independent of each other to be able to fuse them correctly in a Bayesian framework of a forensic AVC system. Some attempts in the research on automatic systems have been made, where some were more successful than others. The Super-SID project[33] at MIT yielded a system using, for example, intonation modelled as bigrams (Reynolds et al., 2003) and phone duration models to improve their automatic system (Adami & Hermansky, 2003). The fused system (i.e. the combination of an automatic system and linguistic information) yielded better results, where the optimal system gave an EER of 0.22% (Reynolds et al., 2003). This project shows that there are ways to improve systems using high-level (i.e. linguistic) knowledge even though it is not clear how or what units to use to make the systems optimal. When it comes to a forensic integration, there is so much information and variation to be handled and modelled that it would be unwise to presume that one can be a completely ultra-objective bystander feeding a system with the necessary inputs to decide the strength of the evidence. However, this does not mean that one should not strive to find the optimal features and measures and find out how to extract them in the most robust manner. Currently a system used for forensic purposes should not include phonetic or linguistic input because the degree of dependence or independence of different components is not known, thus risking undue weight to be introduced to elements of the system. Speaking parameters can then be treated separately through perceptual analysis and be combined at a later stage of analysis. Voice parameters on the other hand can provide useful information in forensic cases and if a

[32] The log energy coefficient is discarded when building the model.
[33] Exploiting High-level Information for High-performance Speaker Recognition.

system attempts to quantify the differences between relevant speakers in a population, it can be a valuable tool.

One important development in bringing together perceptual speech parameters and automatically analysed voice parameters occurred in 2010 at the NIST evaluation with the inclusion of Human Assisted Speaker Recognition (HASR) (Greenberg et al., 2010). Massachusetts Institute of Technology (MIT) Lincoln Labs and United States Secret Service (USSS) jointly developed a phonetic expert system in conjunction with an automatic system called Super Phonetic Annotation and Analysis Tool (SPAAT), which showed very promising results using 15 of the most difficult trials in the bigger NIST corpus (Schwartz et al., 2011a). Their results and methodology encourage more research on the integration and weighting between strengths and weaknesses from the two approaches.

5.4.5 Plugin and Database

During the research reported here, several tools and databases were developed. The general aim has been to use open-source software as far as possible. This was done for reproducibility purposes, but also to be able to understand underlying processes and to be able to adjust and develop the tools needed for the purposes of the thesis, i.e. both for research and practical casework. The tools are described thoroughly in each study; however, a plugin developed for more general use is described here together with a description of the database used for the development of the universal background model.

5.4.5.1 Praat Plugin developed for the thesis

The Praat program is an open-source software (GPL[34]) for phonetic and acoustic analyses (Boersma & Weenink, 2007). The open-source package frequently releases stable versions for Windows, MacOsX and Linux. It has modules for several different research purposes, including speech analysis, labelling and segmentation, speech synthesis, listening experiments, speech manipulation, learning algorithms, statistics and graphics, and it is all highly programmable through its own scripting language[35] or through a subroutine called sendpraat.[36] An easy way to implement a structure and framework is to use the plugin mechanism for the scripting language.

[34] http://www.gnu.org/licenses/gpl.html
[35] http://www.fon.hum.uva.nl/praat/manual/Scripting.html
[36] http://www.fon.hum.uva.nl/praat/sendpraat.html

The possibilities the Praat program offers are almost inexhaustible. The following plugin was developed to facilitate the research in this thesis.

5.4.5.1.1 The ALIZE AVC plugin

AVC is (as already mentioned) an acronym for Automatic Voice Comparison. Automatic methods are increasingly being used in forensic phonetic casework, but most often in combination with perceptual and acoustic methods. For the studies in this thesis using an automatic system, but also as a tool for visualisation and demonstration, text-independent voice comparison using ALIZE was implemented as a plugin to Praat and first presented in Lindh (2009). The purpose of this tool was to create an implementation that was as easy as possible to use, so that people with phonetic knowledge could use the system, perform research or develop it further for forensic casework.

A UBM-GMM technique (Reynolds et al., 2000) was applied with compiled binaries from the ALIZE toolkit (Bonastre et al., 2008; 2005). A standard setup was made for placing data within the plugin. On the top of the tree structure, several scripts controlling executable binaries, configuration files, data etc. were created with basic button interfaces that show up in a given Praat configuration. The scripts were made according to the different necessary steps that have to be covered to create a test environment for comparing voices and rank the likelihood ratios in the output from the system. First of all, some kind of parameterization had to be made of the recordings at hand. In this first implementation, SPro (Guillaume, 2004) was chosen for parameter extraction as there was already support for this implemented in the toolkit. There are two ways to extract parameters: either you choose a folder with audio files (preferably wav format, although other formats are supported) or you record a sound in Praat directly. If the recording is supposed to be of a user of the system (i.e. using the system for verification purposes), a scroll list with a first option “New User” can be chosen. This function will control the sampling frequency and re-sample if the sample frequency is anything other than 16 kHz (default), and perform a frame selection by excluding silent frames longer than 100 ms before 19 MFCCs are extracted and stored in a parameter file. The parameters are automatically energy normalised before storage. The name of the user is also stored in a list of users for the system. If you want to add more users, you go through the same procedure again. When you are finished, you can choose the next option in the scroll list called “Train Users”. This procedure will check the list of users and then normalise and train the users using a background model (UBM). The individual models are trained using MAP (explained in 5.4.2.1 Modelling voices). This procedure requires that you already have a trained UBM. However, if you do not have one, you can choose the function "Train World" which will take your list of users (if you have not added others to be included solely in the world model) and train one with the default of 512 Gaussian mixture models (GMM). The last option on the scroll list is instead "Recognise User", which will test the recording against all the models trained by the system. A list of raw (not normalised) likelihood ratio scores gives you feedback on how well the recording fitted any of the models. In a commercial or fully-fledged verification system, you would also have to test and decide on a threshold for acceptance or rejection, but that is not the aim of this plugin.

5.4.6 UBM Database

To make any kind of quantitative judgement on the discriminative power of parameters used for forensic phonetics, empirical knowledge about the real world is needed. Databases of different kinds can provide us with statistics and opportunities to investigate how well a hypothesis holds. For evaluating the precision and accuracy of a parameter, the Bayesian framework is excellent (Rose, 2002). For the research carried out for this thesis many different databases have been used. The purpose has not primarily been to attempt to define a measurement of how discriminative every parameter is, but rather a more descriptive approach has been adopted. This is mainly because the data has not been ecologically valid, i.e. it is not casework material with non-contemporary recordings (i.e. multiple recordings of individuals made after some time delay, to replicate the fact that in real cases the evidential and reference samples are often separated by months or even years).

5.4.6.1 SweDia

The database used for training the UBM for the purposes of the thesis is SweDia. This database consists of recorded speech material from 107 Swedish locations. The recordings were made as part of a research project carried out as a joint effort by the departments of linguistics at Umeå, Stockholm and Lund universities. The research project was funded by the Bank of Sweden Tercentenary Fund for the period 1998-2003. The full title of the project was The Phonetics and Phonology of the Swedish Dialects around the Year 2000.

Most of the recordings were made during the summer of 1999. The recording locations were evenly distributed over Sweden and the Swedish-speaking parts of Finland, taking into account both geographical distribution and population density. For each location twelve speakers were recorded, representing two age groups - young adults aged 25-35 and an older generation, aged 55-65, approximately equivalent to the parent generation of the younger speakers. Both age groups consisted of three male and three female speakers. In the figure below the geographical distribution of all the recording locations can be observed.

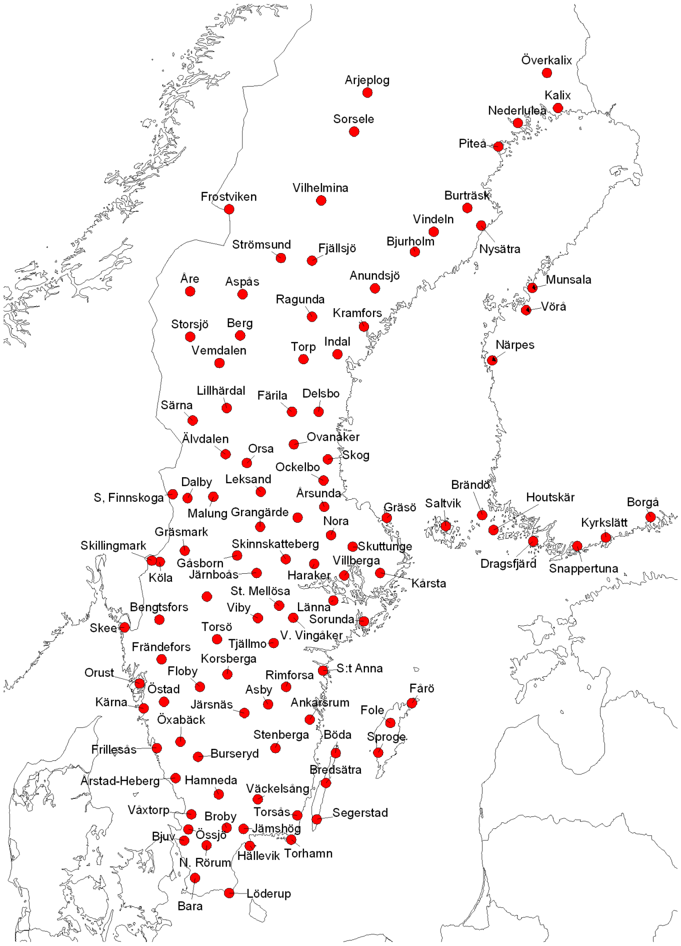


Figure 6. The geographical spread of all places where SweDia recordings took place.

There are other dialect databases with a comparable number of speakers and dialects but this database has certain unique properties:

- Synchronic: All recordings were made within a narrow and very precisely defined time slice. That means that they represent the dialectal variation within the Swedish-speaking community at a precisely defined moment in time.
- Consistent: The recorded material has three parts that contain precisely defined speech material meant to represent three fundamental, phonological properties of Swedish: the quantity system, the accent system, and the phoneme inventory. This means that it is possible to analyse and compare speech material that is identical for all dialects.

- Complete: In addition, the database contains approximately 30 minutes of spontaneous speech per speaker, corresponding to 80-100,000 running words per dialect when transcribed. This part provides information about how the various phonological rules are implemented in everyday casual speech.

The database may also be used for morphological and syntactic studies. These different, well-controlled components make the database a unique source of information for answering research (or case) questions of typicality, i.e. the breadth of distribution of different parameters in a varied and detailed way.

There are several fundamental scientific questions that may be addressed in a fruitful way by an analysis based on data of the kind contained in the SweDia database. One such question which has played a central role in the SweDia project will be described in some detail. In traditional dialectology, geographical dispersion and isolation of groups of speakers as well as renewed contact as a result of migration are considered the driving forces behind linguistic variation and change. There is no doubt that those factors are important, but is that all there is? If it were, then variation and change could take any direction and the end result would appear more or less chaotic. That is not the case, however. Linguistic laws must exist which govern development and change, constrain variation and make certain choices more likely than others. A basic tenet in the SweDia project was the belief that those laws and constraints can be discovered, described and tested. Interesting results have been obtained by approaching the description of regional distribution from an angle that does not assume any particular geographically based constraints at all. In three studies, Leinonen (2009), Lundberg (2005) and Schaeffler (2005), cluster analysis has been used as a means of constructing dialect "areas" based only on individual acoustically grounded phonological properties. The resulting geographical areas are defined by dialects whose properties cluster together. In principle this could result in a very scattered picture with no obvious geographical coherence. But this does not turn out to be the case. On the contrary, dialects group in geographical areas that in many cases closely resemble those suggested in traditional dialectology. This would be a rather trivial finding if the clustering was based on the same considerations as the traditional analyses, but this is not the case at all. In both studies mentioned above, the cluster analyses are based on acoustic properties never considered in traditional classifications. This lends support for the assumption that dialectal variation is constrained by the compatibility of internal factors. Dialectal variation as a function of age and gender of the speaker may also be studied using the

SweDia database. Previous studies have shown that these factors play a role, but the size of the SweDia database and the fact that the material is identically structured for all speakers represented in the database makes it possible to address this question with a greater degree of precision than has previously been possible.

Another important area of research is typologically oriented studies of sound systems and accent systems. Many examples of such studies can be found by looking at the SweDia publication list[37] but many more remain to be carried out.

In order to make the speech data searchable, the audio files have partly been tagged (time aligned annotation). The tagging consists of text files containing various relevant labels and information about where in the sound file the corresponding feature may be found. There are several analysis programs available which allow the audio files to be displayed (and listened to) with the tag files time aligned with the audio signal in a separate window. The sound format and a tag file format are compatible with the two most widely used (freeware) analysis programs, Praat and WaveSurfer (Lindh & Eriksson, 2009). The database in its current format needs at least 200 GB of disk space. The tools for acoustic research can be used in the open-source environment Praat.

To increase the training material for the Universal Background Models (UBM) used here, it is important to have access to databases with data representative for the population in question. For that purpose, 2 UBMs (1 male and 1 female) were trained on the data from the SweDia database. For this an automatic extraction method was created. The program automatically trawls through the database and extracts two minutes of spontaneous speech from each speaker, which is then used for the UBM training. 109 locations x 3 speakers x 4 categories (young and old males and females) = 1308 speakers and 2616 minutes in total. For the studies in this thesis only the UBM based on speech from 628 male speakers were used.

[37] http://www.ling.gu.se/~anders/SWEDIA/publ_eng.html

6 GENERAL CONTRIBUTIONS OF THE THESIS

The work has shown that there can be a difference between what can be considered to be voice and what can be considered speech, contributing to new ways of looking at this kind of forensic comparison. A major contribution is the clarification and investigative research on the robustness and distinctiveness of fundamental frequency measures. The thesis is an attempt to build a bridge between two separate techniques and research areas, but maybe most importantly to demonstrate how to investigate them and apply them to real-world casework. Building the bridge allows the possibility of more and new research on the relation between perceptual human judgement and statistical, acoustically based methods (often referred to as biometric). An understanding of the differences between the two techniques is needed, partly to be able to apply them, but also to be able to explain them when undertaking casework and presenting evidence in court. Looking back, the cooperation between the linguistic phonetic discipline and the engineering community working on automatic acoustic methods has been very sparse. With one foot in each discipline, the hope of this work is to take a leap towards more intense and important cooperative research. The work done here attempts to clarify strengths and weaknesses of the two approaches and show how the two can be used in conjunction with each other. Even though a lot more work has to be done on how to combine them, maybe the most important thing is to know how to deal with the conflict between them.

6.1 Summary of Findings

6.1.1 Study I

The idea that inspired Study I was that a model based on the principles underlying the Modulation Theory of Speech (Traunmüller, 1994) could be used to construct a fundamental frequency measure that better represents the neutral fundamental frequency level of a given voice and at the same time would be more robust to variations in speaking style and channel distortions. The hypothesis was that the alternative fundamental frequency baseline is such a measure. The results from the study lend support to this hypothesis. The results showed that the effects of channel distortion in the speech samples were almost completely neutralised by using the alternative baseline approach. In the experiment where vocal effort was varied, however, a constant fundamental frequency is not to be expected. In this case, the alternative baseline approach still gave a better and more detailed picture of the effect of variation in vocal effort level on fundamental frequency. It is clear, however, that raising vocal effort level will readjust settings of phonation and increase the baseline level. It is thus possible to conclude that for all conditions tested here the alternative baseline method seems to be the most reliable, and that the remaining errors in general seem minimal. In contrast, using the mean or the median often produced substantial errors. The mean shows the highest standard deviation, the median slightly less and the alternative baseline even less. Modelling creakiness must, however, be done separately as it presents a potential problem when the amount of creakiness affects the percentage of values used to extract the alternative baseline. Possibly using the alternative baseline as a point of origin of voice can help further research on modelling other features connected to prosody.

6.1.2 Study II

Study II explores the similarities between an ear-witness line-up experiment and the results from two listening tests (one controlled and one uncontrolled web based test), where subjects judged the voice similarity level between the recorded participants for the ear-witness study. Besides comparing the results, the aim was to see whether some basic speaking tempo parameters (articulation rate and pausing) could explain the results. The ear-witness study (Öhman, Eriksson & Granhag, 2008) and the listening tests of voice similarity showed the same bias between false acceptance and mean voice similarity judgments in respect of three voices used in the line-up. The same three voices were also the ones that were closest to the target voice in terms of pausing and articulation rate. The voice that was most often confused with the target voice in the line-up did not only show the highest perceptually judged voice similarity and closest pausing and articulation rate, but also the highest mean rank compared to the other voices in voice similarity judgments. This indicates that the combination of being a so-called wolf voice (a good imitator or highly confusing general voice) (Doddington, Liggett, Martin, Przybocki & Reynolds, 1998) and being similar and very unremarkable in terms of pausing and articulation rate is a case to be aware of in perceptual judgments of speaker similarity.

6.1.3 Study III

For Study III, the automatic system described in ALIZE - an open-source platform for biometrics - was trained and used. The same material as described in Study II was used as test material for the automatic voice comparison system. The results were then compared to the listening test on judging voice similarity also described in Study II[38]. Rankings of similarity were then used to compare the results. Interestingly, the results showed that the same speaker described in Study II also showed the highest mean rank for the automatic system test. The test performed with the automatic system showed that the same voice was also sensitive to impostors, i.e. a so-called lamb voice (Campbell, 1997). An attempt was then made to use multidimensional scaling to cluster the different voices in groups according to the scores from the automatic system and rankings from the listening test. A similar approach was taken in McDougall (2013) to test perceptual voice similarity for finding a proper set of foils for voice parades. The multidimensional scaling showed that several speakers clustered together similarly for both the perceptual test and the automatic systems. One group seemed to be the speakers closest to the target voice and had the highest overall scores. The other group seemed to be a group of speakers that were almost never confused with others, neither in the listening test nor in the automatic system test. This study shows that there seems to be some correlation between what listeners perceive as similar voices and what an automatic system bases its scores on. There are also indications that when the listening task is rather difficult, i.e. containing short recordings of a group of homogeneous speakers, where they share some basic features such as articulation rate and pausing and the voice quality is rather general, the speakers are easily confused.

[38] The author acknowledges that study III repeats some of the text of study II, in order to describe the material and listening test referred to in both studies. Study III develops Study II by comparing the results of Study II with those of an automatic system. The two papers are proceedings articles from the same conference, FONETIK held in 2009 and 2010. The research was conducted at the time for a standard research-based PhD rather than the current publications-based PhD. The repeated text for the two studies was derived from that ongoing PhD thesis, and was used in Study III to develop new lines of enquiry and argumentation. The text was therefore “recycled” in the sense described by Hexham (2013) as a common process for academics converting their PhDs into published works. Unfortunately Study II was not cited in Study III.

6.1.4 Study IV

In Study IV, a proposal for how to compare the performance between humans and automatic systems is suggested. As a part of the NIST Speaker Recognition Evaluation (SRE) 2010, a Human Assisted Speaker Recognition test was given (Greenberg et al., 2010). For that test a need for a practical way to compare human performance as well as performance by an automatic system is needed. The same data from the listening test described in Study II and Study III was used. A procedure for comparing the listeners' performance using the test samples played backwards was compared to their performance for the same recordings played forwards. In the next step, the results are compared to the result from the automatic system. The procedure shows how it is possible to convert the listeners' judgments to scores similar to the scores produced by the machine. The scores can go through the same calibration procedure as the machine. The results are then presented using Tippett plots and C_{llr}. Using this small dataset from the ear-witness experiment described in Study II, the automatic system outperformed the listeners. Different explanations for that are presented in Study III. However, the crucial outcome of this study is not that result, but rather the demonstration of a valid procedure for actually making the comparisons.

6.1.5 Study V

In the more extensive Study V, a real forensic case of disputed utterance analysis is presented. The case consisted of a tape recording of a witness to a murder with several undisputed samples of the words in question. A procedure for using relevant data, quantitative measurements and statistical models for calculating likelihood ratios is presented. Due to the limited amount of data often present in forensic cases, methods for handling small datasets with uncertain population distributions are also presented. Repeated acoustic measurements of Voice Onset Time (VOT) and formant frequencies one and two (F1 and F2) were used as data for the analyses. In the original case report, only visual representations of the measurements were used as justification for the subjective conclusion on an ordinal scale (Nordgaard, Ansell, Drotz & Jaeger, 2011) with a verbal expression of strength. In the study, an actual calculation of a likelihood ratio supports the original conclusion. A first attempt was made using a single Gaussian model; however, since the VOT data is not normally distributed, a transformation for that data was applied before the likelihood ratio calculation using a single multivariate Gaussian and Probability Density Function (pdf) for each model could be applied. The obtained LR was extremely high (2 x 10^{77}) which cannot be accepted as realistic given the limited data in the case.

The limitations of the modelling procedure are discussed in the study as well as a methodology on reliability evaluation. The reliability evaluation uses Gaussian distributions of Monte Carlo simulations randomly generated from the mean vectors of the undisputed sample data. The likelihood ratio of the disputed utterance was then calculated ten thousand times using ten thousand sample sets of the simulated models based on the original undisputed data. Using this kind of reliability test for the Gaussian modelling approach still generated extremely high values (99% confidence at least 2 x 10^{44}). This means that even though the reliability method seems to be a valid way of checking the result, the simulations are still based on the original limited data means. If those means are still based on very limited data, our results are not valid.

In a second approach, a Hotelling's T^2 modelling technique was applied instead (Hotelling, 1992). This technique is a standard approach for taking into account the limited data and has been used previously in forensic research (Curran, Triggs, Almirall & Buckleton, 1997a; Curran, Triggs, Almirall, Buckleton & Walsh, 1997b). The technique was applied using two sample statistics where the second sample mean is the disputed utterance data (i.e. sample size one) together with each undisputed sample mean. The likelihood ratio is then calculated as the ratio between the likelihood of the disputed sample plus the undisputed prosecution hypothesis data and the disputed sample plus the undisputed defence hypothesis data (again using pdfs). This was then also done using the same simulations as in the first approach. Based on these calculations a much more moderate likelihood ratio was reached (at least 2 x 10^{8} with 99% confidence).

A third approach was applied which is a Bayesian approach previously demonstrated to be useful for Automatic Speaker Recognition (ASR), Posterior predictive density (Villalba & Brümmer, 2011). It is a method of integrating out nuisance parameters using other external distributions. For this approach prior distributions for VOT were found in (Lundeborg, Larsson, Wiman & McAllister, 2012) and formant measurements were taken from a subset of 202 Swedish females from the middle region in the Swedia database (Lindh & Eriksson, 2009). The generated likelihood ratios using this approach yielded were close to the LRs produced by the second approach (2 x 10^{11}). A last test was also made using "uninformed" priors of zero for every mean in each prior vector. The likelihood ratio obtained for that test was considerably smaller (3 x 10^6).

The study as a whole shows several possible ways to approach calculation of likelihood ratios in a disputed utterance case yielding on a verbal LR scale more or less the same result as the original case report, i.e. acoustic measurements of the disputed utterance are very much more likely given the prosecution hypothesis than the defence hypothesis.

7 GENERAL CONCLUSIONS

This section provides a brief summary of the overall conclusions of the research. The research questions for the thesis were as follows:

Objective A and study I:

RQ1. What could actually cause the sometimes substantial intra-variation for fundamental frequency and how would it be possible to model it or ignore the redundant variation?

RQ2. Which long-term measure of fundamental frequency would be the most robust and usable in the forensic context?

Objective B and studies II, III and IV:

RQ3. What can be predicted by comparing voice similarity for listeners in an ear-witness line-up compared to listeners' voice similarity judgments?

RQ4. Do the same mistakes occur with automatic systems as with listeners' judgments of voice similarity, and why or why not?

RQ5. Is it possible to compare the results of listeners' judgments of voice similarity and scores from an automatic voice comparison system?

RQ6. Is the AVC system better at discriminating between voices than human listeners when speech feature similarities are in conflict with voice similarities?

Objective C and study V:

RQ7. How can a likelihood ratio be calculated for acoustic measurements from disputed utterances in a forensic analysis with very little reference data available?

7.1 Objective A

In the thesis the comparison of voices, speakers and speech (in the form of disputed utterance) have been investigated in the context of forensic phonetics. Study I of the parameter F0 concludes that using an alternative baseline as a measure will diminish the effect of low-quality recordings or varying speaking liveliness producing an answer to RQ2. However, both extremely creaky voices and raised vocal effort induce intra-variation problems that are not solved through this study, although the result encourages new research answering RQ1.

7.2 Objective B

Study II was done in conjunction with a larger project investigating the performance of ear witnesses in three studies (Öhman, Eriksson & Granhag, 2013; Öhman et al, 2008; 2010). It was discovered that humans seem to be much more focused on similarities of speech style and also have problems separating *voice* and *speech* features even when speech is played backwards. The results show that the result from a line-up study can be predicted by perceptual pairwise judgment of voice similarity, answering RQ3.

Study III investigates the differences between an automatic voice comparison system and humans' perceptual judgments of voice similarity. When speech feature similarities (pausing and articulation rate) deviated from voice similarities and there are so-called wolf or lamb voices (or models) involved, listeners were biased toward using the speech similarities. This is a reminder that some voices in some circumstances are maybe so "general" or "average" that in a forensic comparison case, they cannot be distinguished as more or less likely to support any hypothesis. This, however, does not mean that one should ignore speech features as parameters when comparing speakers, just that one should separate them from the feature of voice. From the results of the experiments, it is possible to see a correlation between how speakers were judged as more or less different using multidimensional scaling of similarity ranks compared to both the automatic system and the listeners. However, there are also differences due to the fact that human listeners include information about speech style and have difficulties weighting the parameters, i.e. ignoring them when they are contradictory. Since phonetic features such as speaking tempo correlated well with listeners' judgments even when samples were played backwards it was difficult to see whether it would be possible to correlate any kind of perceptual voice-similarity judgments with the judgments made by the machine. So, to answer RQ4 regarding the mistakes made by humans and the automatic system, there might be a correlation between voice similarity judgments and the machine. However, in general, when including speaker features, the two are not compatible and should therefore be treated differently. The research does not therefore answer this question fully.

Possible ways to compare the different methods in the likelihood ratio framework have been presented successfully in Study IV. The study presents a new functional method for how to convert the similarity judgments made by humans and then compare those to the system results thereby answering RQ5. It was also discovered that the automatic system outperformed the naive human listeners in this task using the small dataset, answering RQ6 to some extent. However, it remains to be tested whether human experts in phonetics and voice can perceptually discriminate between what is *voice* and what is *speech* when trained to do so. It would be beneficial to compare forensic phonetic experts to automatic systems using this methodology and some research has started doing that for forensic evaluations. An AVC system can assist the phonetic expert in the assessment of the importance of the voice parameter.

7.3 Objective C

Study V has investigated several statistical modelling techniques and has shown that it is possible to calculate likelihood ratios using simulations based on existing reference data, answering RQ7. The study also presented several problems with modelling small datasets and developed methods which take into account the lack of data.

8 APPLICATIONS AND FUTURE RESEARCH

8.1 Applications in casework

This thesis sought to build a bridge between forensic phonetics and automatic voice recognition. It also had the important aim of improving real-world casework. The practical casework implications have been considered throughout the work conducted for the thesis on the basis of my own experience as a forensic caseworker and through collaborative interaction with other parties working in the field, both in research and also in forensic practice and law enforcement.

I have conducted casework since 2005, and have been involved in more than 400 forensic cases and given testimony in several countries. Much of this casework has been in cooperation with the Swedish National Forensic Centre (NFC), which has been important in understanding their procedures and their experience of working with expert evidence and its presentation both to clients and the judicial system. Other casework and consultancy has been conducted for SÄPO (Swedish Security Service), FBI, NFI (Netherlands Forensic Institute), NBI (National Bureau of Investigation, Finland) and Norwegian Police. I am also a member of the executive committee of IAFPA.

The thesis has value for practical casework in the following ways.

8.1.1 The use of a robust F0 parameter in casework – Study I

The first study sought to test the different F0 measures currently used in forensic casework by different labs. After studying the literature both in a forensic context as well as in a theoretical linguistic or phonetic context, the most robust measure was the alternative baseline found in study I. However, this parameter would only produce valuable information in exceptional cases, i.e. where there are extreme values relative to the normally distributed values in a population (Lindh, 2006). The conclusions drawn give a clue as to what is possible today and how useful this feature is in a casework context, i.e. in most cases not very useful. The suggested alternative baseline is the most robust measure to use, but at the same time not very robust in the presence of creakiness. Handling creaky phonation could be a subject of future studies.

8.1.2 Combination of AVC and Perceptual Analysis – Studies II, III, IV

In casework, it has been a struggle to find reasonable ways to interpret and combine the results from an automatic system with other linguistic and phonetic parameters and features currently assessed by subjective perceptual judgments. At my lab nowadays most cases involve the use of two commercial AVC systems (Batvox 4.1 and iVocalise). The results from the automatic systems provide one important clue to the full conclusion in most cases. This result is considered to be a component in the analysis of voice similarity, and is treated as independent of other parameters that are also used in casework, such as fundamental frequency (F0), articulation rate and vowel formant frequencies (if examples exist in a similar context and are measurable/available and in matched acoustic conditions). Studies II, III and IV confirm this approach as the most appropriate although it is in principle possible to combine results from AVC and phonetic linguistic analyses. However, combining them requires complex mathematical computations to take into account the correlations between parameters, which we are only just beginning to understand (Gold & Hughes, 2014).

For the AVC components one can test the robustness of a system through system evaluations. At my lab, a parallel has been developed for the perceptual phonetic components, in part based on a similar protocol used by the NFI (Cambier-Langeveld, 2007). In casework, a blind test is regularly created, where an analyst holistically describes all similarities and differences between the samples presented. The purpose is to group similar voice samples from an unknown set containing foils to avoid bias and thus provide an independent test of the perceptual analysis.

8.1.3 Cases of Disputed Utterances Analysis – Study V

A very interesting discovery related to study V regarding cases of disputed utterances is that it is possible to calculate a robust LR based on a very small dataset as confirmed by simulations of much larger datasets. Whether this is a fluke because of the nature of this particular case remains to be seen. However, if replicable, this methodology could potentially be applied to other forensic disciplines. This would be a very rewarding outcome as the number of cases meeting the criteria of well-defined hypotheses as in the example case is very rare. However, the encouraging results motivate much more research in this area.

8.2 Outlook

Finally, the research approach and practical implications discussed in this thesis must be considered in the current practical and research context. This is especially essential given the international development of the accreditation protocol ISO17025 for forensic analyses (Drygajlo et al., 2015; Meuwly et al., 2016). There is clearly a growing consensus that the most productive way forward for the field is the continued collaboration between engineers and phoneticians. The dominant theme of research in forensic phonetics is to explore and improve corpus-based analysis of both speech and voice parameters in a Bayesian LR framework. However, the use of such a framework faces a major obstacle in some countries, including the UK, after a court explicitly rejected the use of LR based conclusions (Berger, Buckleton, Champod, Evett & Jackson, 2011). In Sweden this is not a problematic issue since there is no admissibility criterion, but instead a principle of “free evidence” (“fri bevisföring”). Moreover, there is a standard scale in use for all types of forensic evidence produced by the NFC, which is in effect a verbal likelihood scale (Champod & Evett, 2000; Champod & Meuwly, 2000; Nordgaard et al., 2011). Specifically it is a 9-point ordinal scale from +4 to -4, centred on 0. Despite this there remains a question over how well the court understands the strength of evidence and its limitations just as there is elsewhere (Berger et al., 2011; Cudmore, 2011; Robertson, Vignaux & Berger, 2016; Sjerps & Berger, 2012).

A further consideration for the future is the specific ways in which engineers and phoneticians might best cooperate. A number of initiatives have been taken such as the Interspeech special event on “Speaker comparison for forensic and investigative applications”, the ENFSI monopoly project (Drygajlo et al., 2015), and the explicit welcome extended to scholars in AVC at the IAFPA annual meetings. However, in the best interests of the forensic and judicial context, even wider national and international projects and gatherings are required in order to share resources, skills, knowledge and best practice. The success of future collaboration also requires an open mind from the engineering world in understanding the linguistic complexity of spoken language, and an open mind from phoneticians in understanding the potential of using automatic systems to solve some of the major issues we are facing in forensic analyses. The national effort GSLT (Graduate School of Language Technology[39]), which this PhD work is associated to, is an excellent example of how engineers and linguists have come together to create an extremely rewarding community of researchers with a leg in both disciplines.

[39] http://www.gslt.hum.gu.se/

9 REFERENCES

Adami, A. G., & Hermansky, H. (2003). Segmentation of Speech for Speaker and Language Recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003).* (pp. 841–844). Geneva, Switzerland: ISCA.

Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester, UK: John Wiley.

Aitken, C., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *53*(1), 109–122.

Alderman, T. (2004). The use of Australian-English vowel formant data sets in forensic speaker identification. In *Proc. 10th Australian Intl. Conf. on Speech Science and Technology* (pp. 177-182).

Alderman, T. (2005). *Forensic Speaker Identification: A Likelihood Ratio-based Approach Using Vowel Formants*. Europe, Munich: LINCOM.

Alexander, A. (2005). *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched condition*. PhD Thesis. Ecole Polytechnique Federale De Lausanne.

Alexander, A., Botti, F. and Drygajlo, A. (2004) Handling mismatch in corpus-based forensic speaker recognition. Proceedings of 2004: A Speaker Odyssey, Toledo, Spain, 69–74.

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech Language and the Law*, *12*(2), 174-213.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, *52*, 1687–1697.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, *55*, 1304–1312.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, *64*(4), 460–475.

Bachorowski, J.-A., & Owren, M. J. (2001). Not all laughs are alike: voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science*, *12*, 252–257.

Bachorowski, J.-A., & Smrkovski, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, *110*, 1581–1597.

Baldwin, J., & French, P. (1990). Forensic Phonetics. In J. Baldwin & P. French (Eds.) (pp. 42–64). London: Pinter.

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models. In *Proceedings of Interspeech* (pp. 1505–1508).

Berger, C. E., Buckleton, J., Champod, C., Evett, I. W., & Jackson, G. (2011). Evidence evaluation: a response to the court of appeal judgment in R v T. *Science & Justice*, *51*(2), 43-49.

Bernard, J. (1970). Towards the acoustic specification of Australian English. *Zeitschrift Für Phonetik*, *2*(3), 113–128.

Black, J. W., Lashbrook, W., Nash, E., Oyer, H., Pedrey, C., Tosi, O. I., & Truby, H. (1973). Reply to “Speaker identification by speech spectrograms: some further observations.” *The Journal of the Acoustical Society of America*, *54*, 535–537.

Boë, L.-J. (2000). Forensic voice identification in France. *Speech Communication*, *31*, 205–224.

Boë, L.-J., Bimbot, F., Bonastre, J.-F., & Dupont, P. (1999). Des évaluations des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Languedoc Medical*, *2*(4), 270–288.

Boersma, P., & Weenink, D. (2007). *Praat: doing phonetics by computer (Version 4.5.18) [Computer program]*. Institute of Phonetic Sciences, University of Amsterdam. Retrieved from http://www.praat.org/

Bogert, B. P., Healy, M. J. R., & Tukey, J. W. (1963). The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis* (pp. 209–243).

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1969). Identification of a speaker by speech spectrograms. *Science*, *166*(3903), 338-342.

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1973). Speaker identification by speech spectrograms: some further observations. *The Journal of the Acoustical Society of America*, *54*, 531–534.

Bonastre, J.-F., Bimbot, F., Boe, L.-J., Campbell, J., Reynolds, D., and Magrin-Chagnolleau, I. Person authentication by voice: a need for caution. In Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003) (Geneva, Switzerland, 2003), pp. 33–36.

Bonastre, J. F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Ewans, N., Fauve, B. & Mason, J. S. (2008). ALIZE/spkdet: a state-of-the-art open source software for speaker recognition. In Proc. *Odyssey: the Speaker and Language Recognition Workshop.* (p. 20).

Bonastre, J. F., Wils, F., & Meignier, S. (2005, March). ALIZE, a free toolkit for speaker recognition. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on* (Vol. 1, pp. 737-740). IEEE.

Borin, L., Brandt, M.D., Edlund, J., Lindh, J., and Parkvall, M. (2012). Svenska språket i den digitala tidsåldern – The Swedish Language in the Digital Age. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Boss, D. (1996). The problem of F0 and real-life speaker identification: A case study. *Forensic Linguistics*, *3*, 155–159.

Bousquet, P.-M., Matrouf, D., & Bonastre, J.-F. (2011). Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition. In *INTERSPEECH* (pp. 485–488). Retrieved from http://mistral.univ-avignon.fr/doc/publi/11_Interspeech_Bousquet.pdf

Braun, A. (1995). Fundamental frequency -- how speaker-specific is it? In A. B. & J.-P. Köster (Ed.), *Studies in Forensic Phonetics* (pp. 9–23). Trier: Wissenschaftlicher Verlag Trier.

Braun, A., & Künzel, H. J. (1998). Is forensic speaker identification unethical - or can it be unethical not to do it? *Forensic Linguistics*, *5*, 10–21.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*, 1441–1450.

Broeders, A. P. A. (2001, October). Forensic speech and audio analysis forensic linguistics. In *13th INTERPOL Forensic Science Symposium, Lyon, France* (Vol. 26).

Brown, R. (1981). An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. *Language and Speech*, *24*, 295–310.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2-3), 230–275.

Byrne, C., & Foulkes, P. (2004). The “mobile phone effect” on vowel formants. *The International Journal of Speech, Language and the Law: Forensic Linguistics*, *11*(1), 83-102.

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech, Language & the Law*, *14*(2), 223-243.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, *85*(9), 1437–1462. https://doi.org/10.1109/5.628714

Champod, C., & Evett, I. W. (2000). Commentary on A. P. A. Broeders (1999) `Some observations on the use of probability scales in forensic identification’, Forensic Linguistics, 6(2): 228--41. *The International Journal of Speech, Language and the Law*, *7*, 238–243.

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition.

Speech Communication, *31*, 193–203.

Chollet, G. (1991). About the ethics of speaker identification. In *Proceedings of the Eleventh International Congress of Phonetic Sciences, Aix-en-Provence, France, I* (Vol. 1, pp. 397).

Clermont, F., & Zetterholm, E. (2006). F-pattern Analysis of Professional Imitations of “hallå” in three Swedish Dialects. *Proceedings from Fonetik 2006 Lund, June 7–9, 2006*, pp. 25-28.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*, 207–239.

Compton, A. J. (1963). Effects of filtering and vocal duration upon the identification of speakers aurally. *The Journal of the Acoustical Society of America*, *35*, 1748–1752.

Cudmore, A. (2011). The interpretation of written forensic speaker comparison evidence by potential jurors in the UK. MSc dissertation, University of York.

Curran, J. M., Triggs, C. M., Almirall, J. R., Buckleton, J. S., & Walsh, K. A. J. (1997). The interpretation of elemental composition measurements from forensic glass evidence: II. *Science & Justice*, *37*(4), 245-249.

Curran, J. M., Triggs, C. M., Almirall, J. R., Buckleton, J. S., & Walsh, K. A. J. (1997). The interpretation of elemental composition measurements from forensic glass evidence: I. *Science & Justice: Journal of the Forensic Science Society*, *37*(4), 241–244.

DeCasper, A. J., & Fifer, W. P. (2004). On Human Bonding: Newborns Prefer Their Mothers’ Voices. *Readings on the Development of Children*, 1174–1176.

DeCasper, A. J., & Sigafoos, A. D. (1983). The intrauterine heartbeat: A potent reinforcer for newborns. *Infant Behavior & Development*, *6*(1), 19–25.

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns’ perception of speech sounds. *Infant Behavior & Development*, *9*(2), 133–150.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513–1528.

de Pinto, O., & Hollien, H. (1982). Speaking fundamental frequency characteristics of Australian women: then and now. *Journal of Phonetics*, *10*, 367–375.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). Sheep, Goats, Lambs and Wolves - A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In *International Conference on Spoken Language Processing (ICSLP)* (pp. 1351–1354). Sydney, Australia.

Doddington, G. R. (1985). Speaker recognition---Identifying people by their voices. In *Proceedings of the IEEE* (Vol. 73, pp. 1651–1664). IEEE.

Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A. (2001). The NIST Speaker Recognition Evaluation-Overview, Methodology, Systems, Results, Perspective. *Speech Communication*, *31*, 225–254.

Doherty, E. T., & Hollien, H. (1978). Multiple-factor speaker identification of normal and distorted speech. *Journal of Phonetics*, *6*, 1–8.

Dommelen, W. A. van. (1987). The contribution of speech rhythm and pitch to speaker recognition. *Language and Speech*, *30*, 325–338.

Dommelen, W. A. van. (1990). Acoustic parameters in human speaker recognition. *Language and Speech*, *33*, 259–272.

Dommelen, W. A. van. (1993). Speaker height and weight identification: a re-evaluation of some old data. *Journal of Phonetics*, *21*, 337–341.

Dommelen, W. A. van, & Moxness, B. H. (1995). Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech*, *38*, 267–287.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2015). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic

Speaker Recognition. *European Network of Forensic Science Institutes*. Retrieved from http://www.enfsi.eu/sites/default/files/documents/guidelines_fasr_and_fsasr_0.pdf

Drygajlo, A., Meuwly, D., and Alexander, A. (2003). Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition. In Proc. Eurospeech 2003, pages 689–692, Geneva, Switzerland.

Ellis, S. (1990). `’'It's rather serious...''. In *Early speaker identification*, H. Kniffka (Ed.) (pp. 515–521). Tübingen: Max Niemeyer Verlag.

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America*, *49*, 1842–1848.

Eriksson, A. (2005). Tutorial on forensic speech science. In *Proc. European Conf. Speech Communication and Technology* (pp. 4-8).

Eriksson, A., & Wretling, P. (1997). How flexible is the human voice? - A case study of mimicry. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *EUSROSPEECH ’97* (Vol. 2, pp. 1043–1046). Rhodes, Greece.

Evett, I. W., Scranage, J., & Pinchin, R. (1993). An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. *American Journal of Human Genetics*, *52*(3), 498.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, *31*(HPL-2003-4), 1–38.

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.

Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech, Language and the Law - Forensic Linguistics*, *10*(2), 203–226. https://doi.org/10.1558/sll.2003.10.2.203

Fraser, H., & Stevenson, B. (2014). The Power and Persistence of Contextual Priming: More Risks in Using Police Transcripts to Aid Jurors' Perception of Poor Quality Covert Recordings. *The International Journal of Evidence & Proof*, *18*(3), 205–229. https://doi.org/10.1350/ijep.2014.18.3.453

French, P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics*, *1*, 169–181.

French, P., & 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, J. P., Nolan, F., Foulkes, P., Harrison, P., & 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), 143-152.

Fry, J., & Thelwall, M. (2008). Measuring the Impact of e-Research: Accounting for Disciplinary Differential in Patterns of Diffusion. In *Proceedings of the Fourth International Conference on e-Social Science, University of Manchester* (Vol. 32). Retrieved from http://www.researchgate.net/profile/Mike_Thelwall/publication/228337539_Measuring_the_Impact_of_e-Research_Accounting_for_Disciplinary_Differential_in_Patterns_of_Diffusion/links/0912f51333fabe417e000000.pdf

Furui, S. (1978). Effects of long-term spectral variability on speaker recognition. *The Journal of the Acoustical Society of America*, *64*(S1), 183.

Gannert, T. (2007). *A Speaker Verification System Under The Scope: ALIZE*. Master's Thesis, KTH - Royal Institute of Technology.

Glenn, J. W., & Kleiner, N. (1968). Speaker identification based on nasal phonation. *The Journal of the Acoustical Society of America*, *43*, 368–372.

Gold, E., French, P., & Harrison, P. (2013, June). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, No. 1, p. 060041). ASA.

Gold, E., French, P., & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics Acoustical Society of America*, *19*(1), 060041. https://doi.org/10.1121/1.4800285

Gold, E., & Hughes, V. (2014). Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice*, *54*(4), 292-299.

Goldstein, U. G. (1976). Speaker-identifying features based on formant tracks. *The Journal of the Acoustical Society of America*, *59*, 176–182.

Gomez, P., Alvarez, A., Mazaira, L. M., Fernandez, R., & Rodellar, V. (2007). Estimating the Stability and Dispersion of the Biometric Glottal Fingerprint in Continuous Speech. In *Proceedings of the 4th International Conference on Non-LInear Speech Processing* (pp. 63–66).

Gonzalez, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics*, *32*, 277–287.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, *7*(515-546).

Graddol, D., & Swann, J. (1983). Speaking fundamental frequency: Some physical and social correlates. *Language and Speech*, *26*, 351–366.

Greenberg, C. S., Martin, A. F., Brandschain, L., Campbell, J. P., Cieri, C., Doddington, G. R., & Godfrey, J. J. (2010). Human Assisted Speaker Recognition In NIST SRE10. In *Odyssey* (p. 32).

Greenwald, M. H. (1979). *The effect of decreased frequency bandwidth on speaker identification by aural and spectrographic examination of speech samples* (Master's thesis, Michigan State University. Dept. of Audiology and Speech Sciences).

Greisbach, R. (1999). Estimation of speaker height from formant frequencies. *Forensic Linguistics*, *6*, 265–277.

Grey, G., & Kopp, G. A. (1944). Voiceprint identification. *Bell Telephone Laboratories Report*, 1–14.

Gruber, J. S., & Poza, F. T. (1995). *Voicegram Identification Evidence* (Vol. 54). Lawyers Cooperative Publishing.

Guillaume, G. (2004). *SPro: speech signal processing toolkit*. Software available at http://gforge.inria.fr/projects/spro.

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language & the Law*, *15*(2), pp. 193-218.

Haberman, W., & Fejfar, A. (1976). Automatic identification of personnel through speaker and signature verification—system description and testing. In *Proc. 1976 Carnahan Conf. on Crime Countermeasures* (pp. 23-30).

Hall, M. C. (1975). *Spectrographic Analysis of Interspeaker and Intraspeaker variables of Professional Mimicry*. MA Dissertation, Michigan State University.

Hargreaves, W. A., & Starkweather, J. A. (1963). Recognition of speaker identity. *Language and Speech*, *6*, 63–67.

Harmegnies, B., & Landercy, A. (1988). Intra-speaker variability of the long term speech spectrum. *Speech Communication*, *7*, 81–86.

Harrison, P. (2004). Variability of formant measurements. MSc Dissertation. Department of Language and Linguistic Science, University of York.

Harrison, P. (2013, April). *Making accurate formant measurements: an empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. PhD Thesis. University of York. Retrieved from http://etheses.whiterose.ac.uk/7393/

Hazen, B. (1973). Effects of different phonetic contexts on spectrographic speaker identification. *The Journal of the Acoustical Society of America*, *54*, 650–660.

Hecker, M. H. (1971). Speaker recognition. An interpretive survey of the literature. *ASHA monographs*, *16*, 1.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738–1752.

Hexham, I. (2013). The plague of plagiarism: Academic plagiarism defined. *Irving Hexham (ResearchGate profile). https://goo.gl/vX0DSy*.

Hollien, H. (1974). Peculiar case of “voiceprints.” *The Journal of the Acoustical Society of America*, *56*, 210–213.

Hollien, H. (1977). Status report of “voiceprint” identification in the United States. In J. S. Jackson (Ed.), *International Conference on Crime Countermeasures - Science and Engineering* (Vol. 2, pp. 29–40). https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=56186.

Hollien, H. (1990). *The Acoustics of Crime*. New York: Plenum Press.

Hollien, H., & McGlone, R. E. (1976). The effect of disguise on voiceprint identification. *Nat'l J. Crim. Def.*, *2*, 117.

Horii, Y. (1975). Some statistical characteristics of voice fundamental frequency. *Journal of Speech and Hearing Research*, *18*, 192–201.

Horii, Y., & Ryan, W. J. (1981). Fundamental frequency characteristics and perceived age of adult male speakers. *Folia Phoniatrica et Logopaedica*, *33*(4), 227-233.

Hotelling, H. (1992). The generalization of Student’s ratio. In *Breakthroughs in Statistics* (pp. 54-65). Springer New York.

Houlihan, K. (1977a). Study I. In In Hollien, H., & Hollien, P. (Eds.). (1979). Current Issues in the Phonetic Sciences: Proceedings of the IPS-77 Congress, Miami Beach, Florida, 17–19

December 1977 (Vol. 9). John Benjamins Publishing.

Houlihan, K. (1977b). Study II. In Hollien, H., & Hollien, P. (Eds.). (1979). *Current Issues in the Phonetic Sciences: Proceedings of the IPS-77 Congress, Miami Beach, Florida, 17–19 December 1977* (Vol. 9). John Benjamins Publishing.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR Upper Saddle River, NJ, USA.

Huber, P. J. (2011). *Robust statistics* (pp. 1248-1251). Springer Berlin Heidelberg.

Hughes, V. (2014, October). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. PhD Thesis, University of York. Retrieved from http://etheses.whiterose.ac.uk/8309/

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language & the Law*, *23*(1). pp. 99-132.

Imaizumi, S., & Kiritani, S. (1989). Effects of speaking rate on formant trajectories and interspeaker variations. *Annual Bulletin Research Institute of Logopedics and Phoniatrics*, *23*, 27–37.

Ingram, J. C. L., Prandolini, R., & Ong, S. (1996). Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics*, *3*, 129–145.

Ishihara, S., & Kinoshita, Y. (2008). How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification. *Proceedings of Interspeech 2008, Incorporating SST 2008, Brisbane, Australia*, 1941–1944.

Jassem, W., Steffen-Batog, S., & Czajka, M. (1973). Statistical characteristics short-term average F0 distributions as personal voice features. In W. Jassem (Ed.) (Vol. Speech Analysis and Synthesis, pp. 209–225). Warsaw: Warsaw: Polish Academy of Science.

Jessen, M., A. Alexander and O. Forth (2014). Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. Proceedings of the Audio

Engineering Society 54th International Conference; London, June 12–14, pp. 28–35.

Jessen, M., Enzinger, E., & Jessen, M. (2013). Experiments on long-term formant analysis with gaussian mixture modeling using vocalise. In *Conference of the International Association of Forensic Phonetics and Acoustics, 2013*.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *The International Journal of Speech, Language and the Law*, *12*, 174–203.

Jiang, M. (1996). Fundamental frequency vector for a speaker identification system. *Forensic Linguistics*, *3*, 95–106.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*, 701–714.

Keierleber, J. and Bohan, T., "Ten Years After *Daubert*: The Status of the States," *Journal of Forensic Sciences*, Vol. 50, No. 5, 2005, pp. 1-10, https://doi.org/10.1520/JFS2004241. ISSN 0022-1198

Kersta, L. G. (1962a). Voiceprint identification. *Nature*, *196*, 1253–1257.

Kersta, L. G. (1962b). Voiceprint-identification infallibility. *The Journal of the Acoustical Society of America*, *34*, 1978.

Kimura, T., Ashida, A., & Niyada, K. (2004). Practical speaker-independent voice recognition using segmental features. *Electronics and Communications in Japan = Denki Gakkai Ronbunshi*, *87*, 398–405.

Kinoshita, Y. (2001) Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants, Ph.D. Thesis, the Australian National University.

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *The International Journal of Speech, Language and the Law*, *12*, 235–254.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language & the Law*, *16*(1), pp. 91-111.

Kipper, S., & Todt, D. (2001). Variation of sound parameters affects the evaluation of human laughter. *Behaviour*, *138*(9), 1161-1178.

Kipper, S., & Todt, D. (2003a). Dynamic-acoustic variation causes differences in evaluations of laughter. *Perceptual and Motor Skills*, *96*, 799–809.

Kipper, S., & Todt, D. (2003b). The role of rhythm and pitch in the evaluation of human laughter. *Journal of Nonverbal Behavior*, *27*, 255–272.

Kirkland, J (2003) Forensic Speaker Identification Using Australian English Fucken: A Bayesian Likelihood Ratiobased Auditory and Acoustic Phonetic Investigation. Unpublished Honours Thesis. Australian National University.

Koenig, B. E. (1986a). Review article: Spectrographic voice identification. *Crime Laboratory Digest*, *13*, 105–118.

Koenig, B. E. (1986). Spectrographic voice identification: A forensic survey. *The Journal of the Acoustical Society of America*, *79*(6), 2088-2090.

Kuitert, M., & Boves, L. (1997). Speaker verification with GSM coded telephone speech. In *Proc. Eurospeech'97*, vol. 2, pp. 975-978.

Künzel, H. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica*, *46*, 117–125.

Künzel, H. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics*, *4*, 48–83.

Künzel, H. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, *7*, 149–179.

Künzel, H. J. (1987). *Sprechererkennung: Grundzuge forensicher Sprachverarbeitung*. Heidelberg: Kriminalistik-Verlag.

Künzel, H. J. (1994). Current approaches to forensic speaker recognition. In G. Chollet (Ed.), *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (pp. 135–141). Martigny, Switzerland: ESCA.

Künzel, H. J. (2001). Beware of the "telephone effect": The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, *8*, 80–99.

Künzel, H. J. (2002). Rejoinder to Francis Nolan's "The "telephone effect" on formants": A response. *Forensic Linguistics*, *9*, 83–86.

Künzel, H., Köster, J-P. & Masthoff, H. (1995). The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. *Science & Justice: Journal of the Forensic Science Society*, *35*, 291–295.

LaRiviere, C. (1975). Contributions of fundamental frequency and formant frequencies to speaker identification. *Phonetica*, *31*, 185–197.

Leemann, A., Kolly, M.-J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59–67. https://doi.org/10.1016/j.forsciint.2014.02.019

Leinonen, T. (2009). Classifying Swedish dialects based on vowel pronunciation. In *Workshop of Production, Perception, Attitude., Leuven, Netherlands* (Vol. 51, p. 99).

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 1695-1699). IEEE.

Lindh, J. (2005). Visual acoustic vs. aural perceptual speaker identification in a closed set of disguised voices. In *Proc. The 18th Swedish Phonetics Conference (FONETIK 2005), Göteborg, Sweden* (pp. 17-20).

Lindh, J. (2006). Preliminary Descriptive F0-statistics for Young Male Speakers. In S. Schötz & G. Ambrazaitis (Eds.), *Working Papers 52: Papers from Fonetik 2006*. Lund, Sweden: Department of Linguistics, Lund University, pp. 89-92.

Lindh, J. (2006) 'Preliminary Descriptive F0-statistics for Young Male Speakers' Lund University

Working Papers 52, 89-92.

Lindh, J. (2008). Robustness of Forced Alignment in a Forensic Context. In *Proceedings of IAFPA2008*. Lausanne, Switzerland: IAFPA.

Lindh, J. (2009). A first step towards a text-independent speaker verification Praat plug-in using Mistral/ALIZE tools. In *The XXIInd Swedish Phonetics Conference, Department of Linguistics, Stockholm University* (pp. 194–197).

Lindh, J. (2010). Preliminary Formant Data of the Swedia Dialect Database in a Forensic Phonetic Perspective. In A. Braun (Ed.), *Proceedings of the 19th Annual Conference of the International Association for Forensic Phonetics and Acoustics*. Trier, Germany: IAFPA.

Lindh, J., & Åkesson, J. (2016). Evaluation of Software "Error checks" on the SweEval2016 Corpus for Forensic Speaker Comparison. In *Proceedings of IAFPA-25* (pp. 57–58). York, UK.

Lindh, J., & Eriksson, A. (2007). Robustness of Long Time Measures of Fundamental Frequency. In *Proceedings of INTERSPEECH 2007* (pp. 2025–2028). Antwerp, Belgium.

Lindh, J., & Eriksson, A. (2009). The SweDat Project and Swedia Database for Phonetic and Acoustic Research. In *Proceedings of the 2009 Fifth IEEE International Conference on e-Science* (pp. 45–49). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/e-Science.2009.15

Lindsey, G., & Hirson, A. (1999). Variable robustness of nonstandard /r/ in English: evidence from accent disguise. *International Journal of Speech Language and the Law*, *6*(2), 278-289.

Linville, S. E. (1988). Intraspeaker variability in fundamental frequency stability: An age-related phenomenon? *Journal of the Acoustical Society of America*, 741–745.

Linville, S. E., & Korabic, E. W. (1987). Fundamental frequency stability characteristics of elderly women's voices. *Journal of the Accoustical Society of America*, *81*, 1196–1199.

Lundberg, J. (2005). *Classifying Dialects Using Cluster Analysis*. Master's thesis, Department of

Linguistics, University of Gothenburg.

Lundeborg, I., Larsson, M., Wiman, S., & McAllister, A. M. (2012). Voice onset time in Swedish children and adults. *Logopedics, Phoniatrics, Vocology*, *37*(3), 117–122. https://doi.org/10.3109/14015439.2012.664654

Magrin-Chagnolleau, I., Bonastre, J.-F., & Bimbot, F. (1995). Effect of utterance duration and phonetic content on speaker identification using second order statistical methods. In *EUROSPEECH '95* (Vol. I, pp. 337–340). Madrid, Spain.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance'. In: Proc. Eurospeech '97. Rhodes, Greece, pp. 1895–1898.

Martin, A. F., & Przybocki, M. A. (1999). The NIST 1999 Speaker Recognition Evaluation-An Overview. *Digital Signal Processing*, *10*, 1–18.

Matrouf, D., Scheffer, N., Fauve, B., & Bonastre, J.-F. (2007). A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification. In *Proc. Interspeech* (pp. 1242–1245).

Mayoue, A. (2007). *Reference system based on speech modality ALIZE/LIA-RAL*. Technical report, GET-INT.

McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /aI/. *The International Journal of Speech, Language and the Law: Forensic Linguistics*, *11*(1).

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers. *The International Journal of Speech, Language and the Law*, *13*, 89–126.

McDougall, K. (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, *20*(2), pp. 163-172.

McLaren, M., Matrouf, D., Vogt, R., & Bonastre, J. F. (2008). Combining continuous progressive model adaptation and factor analysis for speaker verification. Proceedings of Interspeech 2008, pp. 857-860.

Meuwly, D. (2003a). Le mythe de « L'empreinte vocale » (I). *Revue Internationale de Criminologie et de Police Technique et Scientifique*, *56*(2), 219–236.

Meuwly, D. (2003b). Le mythe de « L'empreinte vocale » (II). *Revue Internationale de Criminologie et de Police Technique et Scientifique*, *61*(3), 361–374.

Meuwly, D., & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). In *A Speaker Odyssey - The Speaker Recognition Workshop* (pp. 142–150). Crete, Greece.

Meuwly, D., Ramos, D., & Haraksim, R. (2016). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*. In press DOI: https://doi.org/10.1016/j.forsciint.2016.03.048

Moosmüller, S. (1997). Phonological variation in speaker identification. *Forensic Linguistics*, *4*, 29–47.

Moosmüller, S. (2001). The influence of creaky voice on formant frequency changes. *Forensic Linguistics*, *8*, 100–112.

Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice: Journal of the Forensic Science Society*, *49*(4), 298–308.

Morrison, G. S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*, 2387.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model -- universal background model (UBM-GMM). *Speech Communication*, (53), 242–256.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*(3), 91-98.

Morrison, G. S., & Kinoshita, Y. (2008). Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English/o/Formant Trajectories. In *Proceedings of Interspeech* (pp. 1501–1504).

Morrison, G. S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B.K., De Souza, S., Cummins, N. & Chow, D. (2015). Forensic database of voice recordings of 500+ Australian English speakers. *URL http://databases. forensic-voice-comparison.net*

Geoffrey Stewart Morrison, Farhan Hyder Sahito, Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs, Caroline Goemans Dorny, INTERPOL survey of the use of speaker identification by law enforcement agencies, Forensic Science International, Volume 263, June 2016, Pages 92-100, ISSN 0379-0738, http://dx.doi.org/10.1016/j.forsciint.2016.03.044

Mowrer, D. E., LaPointe, L. L., & Case, J. (1987). Analysis of five acoustic correlates of laughter. *Journal of Nonverbal Behavior*, *11*, 191–199.

Mullennix, J. W., Johnson, K., & Topcu, M. (1991). Context effects in the perception of personal information in the speech signal. *The Journal of the Acoustical Society of America*, *89*(4B), pp. 2011.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America*, *68*, 1294–1300.

Nakasone, H., & Beck, S. D. (2001). Forensic automatic speaker recognition. In *A Speaker Odyssey - The Speaker Recognition Workshop* (pp. 139–142). Crete, Greece.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology-Human Perception and Performance*, *24*(3), 756–766.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1990). Texte zu Theorie und Praxis Forensischer Linguistik, Chapt. The limitations of auditory-phonetic speaker identification. *Max Niemeyer Verlag*, *48*, 457-481.

Nolan, F. (2002a). Intonation in speaker identification: an experiment on pitch alignment features. *Forensic Linguistics*, *9*, 1–21.

Nolan, F. (2002b). The “telephone effect” on formants: A response. *Forensic Linguistics*, *9*, 74–82.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *The International Journal of Speech, Language and the Law*, *12*, 144–173.

Nolan, F., & Oh, T. (1996). Identical twins, different voices. *Forensic Linguistics*, *3*, 39–49.

Nordgaard, A., Ansell, R., Drotz, W., & Jaeger, L. (2012). Scale of conclusions for the value of evidence. *Law, Probability & Risk*, *11*, 1-24.

Öhman, L., Eriksson, A., & Granhag, P. A. (2008). Earwitness identification accuracy in children vs. adults. In *The 18th conference of the European Association of Psychology and Law, Maastricht (Netherlands) July 2-5, 2008*.

Öhman, L., Eriksson, A., & Granhag, P. A. (2010). Mobile phone quality vs. Direct quality: How the presentation format affects earwitness identification accuracy. *The European Journal of Psychology Applied to Legal Context*, *2*(2), 161–182.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Enhancing adults' and children’s earwitness memory: Examining three types of interviews. *Psychiatry, Psychology and Law*, *20*(2), 216-229.

Oppenheim, A. V., & Schafer, R. W. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, *21*(5), 95–106.

Owen, T., & McDermott, M. C. (1996). Voice Identification: The Aural/Spectrographic Method, Chapt. 6. *Lawyers & Judges Publishing Company, Inc*, *2*(0), 22.

Owren, M. J., & Bachorowski, J.-A. (2003). Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior*, *27*, 183–200.

Papcun, G. (1988). What do mimics do when they imitate a voice?. *The Journal of the Acoustical Society of America*, *84*(S114), 466-481.

Petrini, K., & Tagliapietra, S. (2008). Cognitive Maturation and the Use of Pitch and Rate Information in Making Similarity Judgments of a Single Talker. *Journal of Speech, Language, and Hearing Research: JSLHR*, *51*(2), 485.

Pilz, C. S. (2006). *U.S. Patent Application No. 11/482,549*.

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *The Journal of the Acoustical Society of America*, *26*, 403–412.

Potter, R. K. (1945). Visible patterns of speech. *Science*, *November*, 463–470.

Poyatos, F. (1993). The many voices of laughter: A new audible-visual paralinguistic approach. *Semiotica*, *93*, 61–81.

Provine, R. R. (1993). Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology: Formerly Zeitschrift Fur Tierpsychologie*, *95*, 291–298.

Provine, R. R., & Yong, Y. L. (1991). Laughter: A stereotyped human vocalization. *Ethology: Formerly Zeitschrift Fur Tierpsychologie*, *89*, 115–124.

Pruzansky, S. (1963). Pattern-matching procedure for automatic talker recognition. *Journal of Acoustic Society of America*, *26*, 403–406.

Psutka, J., Müller, L., & Psutka, J. V. (2001). Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task. In *INTERSPEECH* (pp. 1813-1816).

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of Speech & Hearing Research*, *9*, 273–277.

Ramos, D., Franco-Pedroso, J., & Gonzalez-Rodriguez, J. (2011). Calibration and weight of the evidence by human listeners. the ATVS-UAM submission to NIST human-aided speaker recognition 2010. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 5908-5911). IEEE.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, *288*(5464), 349-351.

Rao, K. S., & Sarkar, S. (2014). *Robust Speaker Recognition in Noisy Environments*. Springer International Publishing. Retrieved from https://books.google.se/books?id=FdYkBAAAQBAJ

Rathborn, H. A., Bull, R. H., & Clifford, B. R. (1981). Voice recognition over the telephones. *Journal of Police Science and Administration*, *9*, 280–284.

Reich, A. R., Moll, K. L., & Curtis, J. F. (1976). Effects of selected vocal disguises upon spectrographic speaker identification. *The Journal of the Acoustical Society of America*, *60*, 919–925.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, *17*(1-2), 91–108. https://doi.org/10.1016/0167-6393(95)00009-D

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In *Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on* (Vol. 4, pp. IV-4072). IEEE.

Reynolds, D. A. (2003). Channel robust speaker verification via feature mapping. In *Proc. ICASSP* (Vol. 2, pp. 53–56).

Reynolds, D., Andrews, W., Campbell, J., Navrtil, J., Peskin, B., Adami, A., Jin, Q., Klusek, D., Abramson, J., Mihaescu, R. & Godfrey, J. (2002). *Exploiting high-level information for high-performance speaker recognition, super SID project final report*. Technical Report, The Center for Language and Speech Processing, 2002. http://www. clsp./jhu. edu/ws2002/groups/supersid.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*, 19–41.

Robertson, B., Vignaux, G. A., & Berger, C. E. (2016). *Interpreting evidence: evaluating forensic* *science in the courtroom*. John Wiley & Sons.

Rose, P. (1999). Differences and distinguishability in the acoustic characteristics of hello in voices of similar-sounding speakers: A forensic phonetic investigation. *Australian Review of Applied Linguistics*, *22*, 1–42.

Rose, P. (2002). *Forensic Speaker Identification*. New York: Taylor & Francis.

Rose, P. (2003). In the legal reference series Expert Evidence, Chapt. *The Technical Comparison of Forensic Speech Samples. Thomson Lawbook Co*, *2*(6), 37.

Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, *20*(2-3), 159–191.

Rose, P. (2010). 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*(4), 316-329.

Rose, P., Kinoshita, Y., & Ishihara, S. (2008). Beyond the long-term mean: Exploring the potential of F0 distribution parameters in traditional forensic speaker recognition. In *Proceedings of the Odyssey Speaker and Language Recognition Workshop* (pp. 329-334).

Rose, P., & Morrison, G. (2009). A response to the UK position statement on forensic speaker comparison. *The international journal of speech, language and the law*, *16*(1), 139-163.

Rose, P., Osanai, T., & 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*, 179–202.

Rothman, H. B. (1977). A perceptual (aural) and spectrographic identification of talkers with similar sounding voices. In J. S. Jackson. (Ed.), *International Conference on Crime Countermeasures - Science and Engineering.* (pp. 37–42). Oxford: University of Oxford.

Ryan, W. J. (1972). Acoustic aspects of the aging voice. *Journal of Gerontology*, *27*, 265–268.

Schaeffler, F. (2005). *Phonological Quantity in Swedish Dialects: Typological aspects, phonetic* *variation and diachronic change.* Doctoral dissertation, Philosophy and Linguistics, Umea University.

Schmidt-Nielsen, A., & Crystal, T. H. (1998). Human vs. machine speaker identification with telephone speech. *ICSLP 98*, *2*, 221–224.

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *The Journal of the Acoustical Society of America*, *43*(1178-1179).

Schwartz, M. F., & Rine, H. E. (1968). Identification of speaker sex from isolated, whispered vowels. *The Journal of the Acoustical Society of America*, *44*, 1736–1737.

Schwartz, R., Campbell, J. P., & Shen, W. (2011b). When to punt on speaker comparison? *The Journal of the Acoustical Society of America*, *130*(4), 2547.

Schwartz, R., Campbell, J. P., Shen, W., Sturim, D. E., Campbell, W. M., Richardson, F. S., Dunn, R.B. & Granville, R. (2011a). USSS-MITLL 2010 human assisted speaker recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 5904–5907). Prague, Czech Republic: IEEE. https://doi.org/10.1109/ICASSP.2011.5947705

Shipp, T., Doherty, E. T., & Hollien, H. (1987). Some fundamental considerations regarding voice identification. *The Journal of the Acoustical Society of America*, *82*, 687–688.

Shipp, T., Qi, Y., Huntley, R., & Hollien, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of Voice: Official Journal of the Voice Foundation*, *6*, 211–2126.

Sjerps, M. J., & Berger, C. E. (2012). How clear is transparent: Reporting expert reasoning in legal cases. *Law, Prob. & Risk*, *11*, 317.

Smrkovski, M., & Bachorowski, J. A. (2003). Antiphonal laughter between friends and strangers. *Cognition & Emotion*, *17*(2), 327-340.

Smrkovski, L. L. (1975). Collaborative study of speaker identification by the voiceprint method. *Journal-Association of Official Analytical Chemists*, *58*(3), 453-456.

Spence, M. J., & DeCasper, A. J. (1987). Prenatal experience with low-frequency maternal voice sounds influences neonatal perception of maternal voice samples. *Infant Behavior & Development*, *10*(2), 133–142.

Stevens, K. N., Williams, C. E., Carbonell, J. R., & Woods, B. (1968). Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *The Journal of the Acoustical Society of America*, *44*, 1596–1607.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185-190.

Stoicheff, M. L. (1981). Speaking fundamental frequency characteristics of nonsmoking female adults. *Journal of Speech & Hearing Research*, *24*, 437–441.

Tiersma, P., & Solan, L. (2002). The Linguist on the Witness Stand: Forensic Linguistics in American Courts. *Language, 78*(2), 221-239. Retrieved from http://www.jstor.org/stable/3086556

Tosi, O. (1979). *Voice Identification: Theory and Legal Applications*. Baltimore: University Park Press.

Tosi, O., & Greenwald, M. (1978). Voice identification by subjective methods of minority group voices. In *6th Meeting of the International Association of Voice Identification.* New Orleans, LA.

Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J., & Nash, E. (1972). Experiment on voice identification. *The Journal of the Acoustical Society of America*, *51*, 2030–2043.

Traunmüller, H. (1994). Conventional, biological, and environmental factors in speech communication: A modulation theory. *Phonetica*, *51*, 170–183.

Traunmüller, H., & Eriksson, A. (1995). The perceptual evaluation of F0-excursions in speech as evidenced in liveliness estimations. *The Journal of the Acoustical Society of America*, *97*, 1905–1915.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*, 3438–3451.

van Leeuwen, D., & Bouten, J. S. (2004). Results of the 2003 NFI-TNO forensic speaker recognition evaluation. In *Proc. Odyssey 2004 Speaker and Language recognition workshop, ISCA* (pp. 75–82).

Villalba, J., & Brümmer, N. (2011). Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance. *Proc. InterSpeech, Florence, Italy*, 505-508.

Wagner, I. (1995). A new jitter-algorithm to quantify hoarseness: An exploratory study. *Forensic Linguistics*, *2*, 18–27.

Wendler, J., Doherty, E. T., & Hollien, H. (1980). Voice classification by means of long-term speech spectra. *Folia Phoniatrica*, *32*, 51–60.

William J., W. A. van D., & Barry, J. (2005). *The Integration of Phonetic Knowledge in Speech Technology*. (W. A. van D. William J. & J. Barry, Eds.). Dordrecht: Dordrecht, Springer.

Williams, U., & Stevens, K. N. (1972). Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America*, *52*, 1238–1250.

Willis, S., Mc Kenna, L., Mc Dermott, S., O'Donnell, G., Barrett, A., Rasmusson, B., Höglund, T., Nordgaard, A., Berger, C., Sjerps, M. & Molina, JJ. (2015). ENFSI guideline for evaluative reporting in forensic science. *European Network of Forensic Science Institutes*.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, *51*, 2044–2056.

Li, X., & Bilmes, J. (2005). Feature pruning for low-power ASR systems in clean and noisy environments. *IEEE Signal Processing Letters*, *12*(7), 489-492.

Young, M. A., & Campbell, R. A. (1967). Effects of context on talker identification. *The Journal of the Acoustical Society of America*, *42*, 1250–1254.

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, *33*(2), 248–248. https://doi.org/10.1121/1.1908630

Appendix - Attached Papers