DIPLOMARBEIT

Titel der Diplomarbeit

Formant trajectories in forensic speaker recognition

Verfasser

Ewald Enzinger

angestrebter akademischer Grad

Magister der Philosophie (Mag. phil.)

Wien, im Dezember 2009

Studienkennzahl lt. Studienblatt: A 328

Studienrichtung lt. Studienblatt: Allgem./Angew. Sprachwissenschaft

Betreuerin: Univ.Doz. Dr. Sylvia Moosmüller

Acknowledgments

I'd like to thank my advisor Sylvia Moosmüller for her grateful support and guidance, especially in regard to the conclusions drawn from the Viennese monophthongisation process and the subsequent interpretation.

I'm deeply grateful to Timo Becker who attentively introduced me to the field of forensic phonetics and automatic speaker identification and provided me with many suggestions and amendments during the completion of the present work. Without his comments and assurance it would have been a much less detailed account of the procedures embedded in this thesis.

I also want to thank the Acoustics Research Institute of the Austrian Academy of Sciences and the Institute of Linguistics at the University of Vienna, who provided me with the necessary resources to conduct my analyses, and especially Werner Deutsch, with whom I had many lengthy discussions about the statistical evaluation of evidence.

My thanks go also to Julia Brandstätter who provided the segmentations and corrected formant parameters for the experiment.

Lastly I want to give thanks to my family and friends who supported me over the course of this work and in fact my whole studies.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In recent years the number of court cases involving speech recordings of suspects as evidence, for example taken from telephone conversations, has seen a substantial increase. Forensic speech evidence is expected to gain even more importance, as speech communication technologies have become ubiquitous. Likewise the role of expert opinion given by forensic phoneticians is sought more often, as it is necessary to specify the degree of identity between the speaker on the given offending recording and the suspect.

The methods used for identifying speakers by their voice must be steadily developed and evolved to satisfy the new requirements and conditions imposed on them. On the one hand these requirements are referring to the legal role in the judicial system that institutes forensic speaker recognition in the pursuit of reaching a verdict. On the other hand these approaches must be tested under all possible technical circumstances that can arise to ensure proper evaluation of speech evidence.

Only recently, biometric systems for speaker identification are adapted and marketed as tools for forensic laboratories and scientists. They enable fully automatic analysis of audio samples and deliver a score or a categorial decision of speaker identity. Their primary advantage is the reduction of time needed for analysis. However, their black box-like functioning bears the risk of misapplication and misjudgement.

In the present work an approach is presented and subsequently evaluated that combines in some sense traditional phonetic analysis and automatic methods to discriminate between speakers. The Bayesian approach is used for evidence evaluation in the light of a coming paradigm shift in the forensic sciences which is driven by many practitioners in the field.

The aim of this thesis is first to investigate the discriminatory potential of the the use of parametric representations of dynamic features of diphthongs in a likelihood ratio approach of evidence evaluation, based on a speech corpus of Viennese German speakers and secondly to evaluate which diphthong offers the best discriminatory power.

Furthermore, the duration aspects of diphthongs in different prosodic positions for the use in this approach are investigated.

1.1 Thesis outline

The thesis is organised as follows: Chapter 2 gives an overview of the field of forensic speaker recognition in order to familiarise the reader with the basic concepts applied when identifying speakers by their voice. Chapter 3 discusses the methods used for the experiment and the evaluation of the results. Chapter 4 presents the experimental setup and the results obtained, using the tools for evaluation outlined in the previous chapter. Chapter 5 discusses the results and gives an interpretation, as well as directions for further research which became apparent while conducting this study.

Chapter 2

Foundations

This chapter provides an overview of the field of forensic speaker recognition as well as methods for the general assessment of evidence in court cases. After defining the basic terminology, the field of forensic speaker recognition is integrated into the larger discipline of phonetics. The different kinds of parameters used for acoustic-phonetic speaker discrimination are discussed with respect to traditional acoustic-phonetic and automatic analysis. Finally, the Bayesian approach is described as a general framework for the evaluation of evidence, followed by a discussion of its implications and the difficulties yet to overcome.

2.1 Forensic Speaker Recognition

Forensic speaker recognition refers to the analysis and comparison of speech recordings with the goal of reaching a decision on the question of speaker identity. The outcome of the analysis is expressed either as a categorial statement or in terms of probabilities.

The methods developed for forensic speaker recognition are primarily applied in court casework that involves evidence in form of speech recordings, e.g. incriminating phone calls like bomb threats.

Forensic speaker recognition involves the use of mainly two different approaches, namely auditory- and acoustic-phonetic analysis. At the Bundeskriminalamt in Germany these approaches are in use for preparing expert reports for police stations and prosecutors since 1985 (Gfroerer 2006:3).

In recent times automatic systems originally designed for biometric applications that assess the similarity between two speakers for the use in commercial environments like physical or information access control have started to be adapted for the use as forensic tools. This trend to include fully automatic analysis of acoustic parameters should on the one hand be seen as beneficial, due to the high standard of technical performance. On the other hand, the uncritical use of automatic speaker identification systems may lead to potentially unexpected errors, especially when the method-specific constraints are neglected or the chain of causality is broken. Phonetic expertise is therefore still needed, especially in order to select segments of speech that can be compared against each other.

2.1.1 Terminology

Forensic speaker recognition can be subdivided into *naive* and *technical* speaker recognition (Nolan 1983:7). The former describes the situation where a layperson without any training in phonetics or hearing sciences derives a judgement of speaker similarity or dissimilarity. This kind of speaker recognition is basic to human perception and can be performed by virtually every hearing person.

The latter term describes the scientific pursuit of performing the task of identifying or discriminating speakers by their voice using forensic-phonetic analysis methods. This is performed by forensic practitioners who usually have received phonetic and linguistic training.

Technical speaker recognition itself can be further sub-categorised. *Speaker verification* refers to deciding whether a claim of identity between two given recordings, one of them known and one unknown, is valid, using a predetermined similarity threshold (Rose 2002:90). This is applied in biometric systems used for physical access control or in telephone based banking applications.

In *Speaker identification* the speech on a recording from an unknown speaker is ascribed to one of a set of known speakers. This constitutes the general forensic case where a speech sample of an offender should be attributed to one of the known suspects. Nolan (1983:9) cites three kinds of tests that can be performed:

- *Closed tests* imply that the unknown speaker is contained in the set of known speakers and thus can be positively identified as one of them.

- *Open tests* on the contrary do not make this assumption which yields an additional potential outcome, namely that none of the known speakers is sufficiently similar to be identified as the unknown speaker.

- *Discrimination tests* deal with the situation where recordings of two speakers are available and it must be decided if there is enough correspondence between the two samples to declare that they originate from the same speaker.

Nolan (1983:9) notes that while 'speaker discrimination most closely resembles speaker verification', it is associated with identification because it faces different circumstantial characteristics. In verification tasks speakers are cooperative and possible impostors will try to imitate the voice of a speaker while in general speaker discrimination co-operation cannot be expected.

2.1.2 Forensics in the context of phonetics and linguistics

Forensic phonetics refers to the application of phonetics and more generally linguistics for forensic-scientific purposes during police investigations and in court cases that handle speech evidence. It applies acoustic and auditory phonetic techniques and methodology to describe the differences between speakers and involves knowledge of properties of languages and dialects into the process of speaker identification.

The field confines itself to the analysis of spoken language and is not concerned with linguistic authorship identification or profiling which is a major task in general forensic linguistics (Broeders 2001, Olsson 2004).

Another area of application except forensic speaker recognition is given during the process of police investigations where it can be used for *voice profiling* (often called *voice analysis*) in which a crude profile of the speaker is created by a trained phonetician based on one or several speech recordings. This profile includes information about sex, age and possible origin up to the countries or areas where he or she has been brought up or has lived.

The application of phonetic and linguistic knowledge in the domain of forensics and specifically speaker identification has been quite controversial. Nolan (1997:746-747) notes that the priority in phonetic research has been on the 'shared linguistic system' between speakers which led to the view that inter-speaker differences were practically noise, 'rather than developing a theory of "speaker space" [...]'. Thus, it was questioned whether phoneticians are in the right position to give their testimonial for the use as evidence in court. Nolan (1997:747) on the other hand states that if phoneticians do not take their responsibility, people much less knowledgeable in the domain of speech would nevertheless be consulted.

2.1.3 The different levels of variability in speech

Speakers exhibit substantial phonetic and phonological variation in their utterances. This simple fact is recognised in the basis of phonology as variation is used to encode distinctions in meaning in the conveyance of information. Structuralist approaches build an abstraction from the continuous nature of the actual physical phenomenon by dividing speech into segmental units, disregarding temporal information, and by the postulation of a set of abstract (possibly feature-bearing) units that act as a system in an individual language, not accounting for the variation in speech exhibited between and within speakers (Keating 1990). The sources of this variability can be presented along the following dimensions.

Between-speaker variation

First there exists considerable variation in utterances of (linguistically) identical words and sounds between different speakers. In the acoustic domain this variation is explained by the acoustic theory of speech production (Fant 1960) along with the source-filter theory which states that the voicing perceived in a vowel is produced by the vibration of the vocal cords and the vocal tract acts as a filter that changes the voicing waveform into a complex periodic waveform. The acoustical difference between the realisations of the same utterance produced by different people can therefore be ascribed to the physiological differences between the vocal tracts of different people. These include the length and condition of the vocal cords themselves as well as the length of the vocal tract.

However, in addition to this base assumption, the fact that people use different settings of the vocal tract during articulation of the same speech segment has to be recognised, too. Thus, speaker identity rests on both the physiological as well as on the behavioural properties.

These aspects constitute the basis of *between-speaker variation* or *inter-speaker variation.*

Within-speaker variation

The fact that speakers themselves exhibit a substantial amount of variation between utterances of linguistically identical material has long been recognised in the acoustic-phonetic literature (Harrington & Cassidy 1999). Rose (2002:10) remarks that '[it] is a phonetic truism that no-one ever says the same thing in exactly the same way'. This is due to the fact that the articulatory organs cannot produce identical settings of the vocal tract at each utterance of the same sound.

The notion of *degrees of freedom* is often used to describe the flexibility of the speech organs while producing speech. They 'may be manipulated at will [. . .] or may be subject to variation due to external factors such as stress, fatigue, health, and so on' (Nolan 2001:2).

These degrees of freedom become relevant in connection with the realisation of abstract phonological units. The received theory is that of *phonetic targets* associated with each phonological unit which have to be achieved in order to convey the signalled information, which is also incorporated into segmental approaches to speech synthesis. These targets are in turn connected up to construct an utterance, which results in co-articulation effects (Keating 1990:454).

Depending amongst other things on extralinguistic circumstances, as for example the emotional state of the speaker, the speaking rate and other factors build up to a phenomenon called *target undershoot*[1] in which the segmental target is not attained during the articulatory transition between units (Rose 2002:233). In consequence of this and the aforementioned co-articulation effects, a quantitative acoustic assessment of the speaker's sound qualities will yield different values for each measurement, aside from the error induced by the measuring instrument and the recording equipment.

Inter-session variability

The concept of *inter-session variability* extends the notion of within-speaker variation and is caused by different linguistic as well as extralinguistic factors that influence speech depending on the speech situation and circumstances.

Speech is highly influenced by the social situation in which it takes place. The choice of register, style and dialect depend on whom we are talking to (or rather

[1] The concept of target undershoot is not undisputed. See Moosmüller (2007a:490) and Moosmüller (2007b:174) for an account of the notion's shortcomings and theoretical problems.

who is possibly listening). Additionally, general voice characteristics can vary a great deal between sessions depending on health, fatigue and other factors. Therefore, it is important to recognise, especially in the forensic context, that speakers can vary in their speech to a large degree and also that this variation can be exhibited in a non-uniform way between two recordings of different sessions which differ in respect to social situation or emotional state.

Variability is not constrained to the kind of diverging realisations of sounds, but also includes idiosyncrasies performed in other linguistic areas, for example the use of characteristic lexemes found in specific dialect regions or aberrant meanings of common words used in close social groups. These facts must be accounted for during the process of speaker discrimination, as the range spanned by the acoustic correlates of speech of one speaker overlaps to some extent with the one of other speakers.

The following section presents a model that tries to account for the range of information conveyed in an utterance.

2.1.4 A 'voice model' for sources of variability

To account for the variability exhibited by speakers, an explicit model of the different mechanisms that convey information, intentionally or otherwise, during an utterance, is needed. Nolan (1983) provides an approach that covers the linguistic as well as the vocal (motoric) mechanism. Figure 2.1 gives an overview of the faculties involved.

McDougall (2005:6) summarises the model as follows.

> In overview, the model explains the multiple types of information conveyed by speech as originating from a speaker's communicative intent which is transmitted via the interaction of the speaker's linguistic mechanism with his or her vocal mechanism. The linguistic mechanism is made up of a number of components which determine a phonetic plan that is implemented by the vocal mechanism to produce a speech signal. The vocal and linguistic mechanisms are each affected by a number of indexical factors also shown in the model.

According to McDougall (2005:8) the communicative intent 'demarcat[es] this component of information in the model as information which the speaker volitionally conveys, while any additional "informative" information comes under the "indexical factors" [...]'.

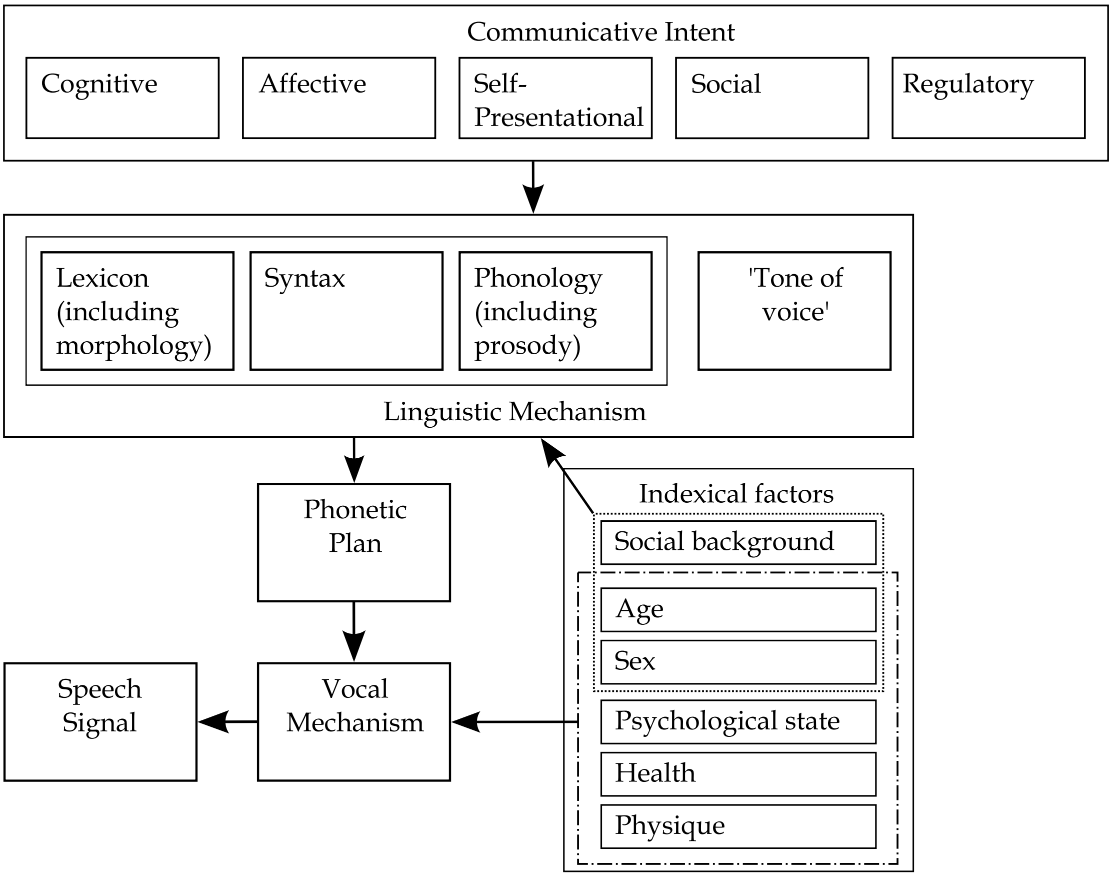


Figure 2.1: Nolan's model of the types of information (McDougall 2005)

The present work makes references to this model at several instances to put in perspective the assumptions underlying different methods for forensic speaker recognition.

2.2 Parameters for speaker discrimination

In order to discriminate between speakers, parameters have to be defined which objectively allow to characterise a speaker. The choice of parameters used in the comparison of speech recordings depends largely on their respective quality and the language of the speech samples to be compared.

As described in the previous section, variation exists between speakers as well as within utterances produced by the same speaker. The logical consequence must therefore be that, to be able to differentiate between speakers, the inter-speaker variation must usually be larger than the intra-speaker variation.

Table 2.2 taken from Rose (2002:34) gives an overview and a rough categorisation of forensic phonetic parameters.

	Linguistic	Non-linguistic
Auditory	**Auditory-Linguistic**	**Auditory-non-linguistic**
Acoustic	**Acoustic-Linguistic**	**Acoustic-non-linguistic**

Table 2.1: Categorisation of forensic-linguistic parameters (Rose 2002:34)

Acoustic vs. auditory parameters

Forensic phonetic parameters can first be categorised along the distinction between auditory and acoustic parameters.

The focus in *auditory analysis* lies on comparing samples with respect to the sound system and language used by the speaker. Initially this procedure involves the task of listening to the speech recording, which should ideally be performed by a trained phonetician, to detect certain cues present in the speech sample which are of use to speaker identification. These characteristics include aspects of voice quality as well as the language variety or dialect used by the speaker. The phonetic segmentation and transcription of the utterance using notations like the International Phonetic Alphabet (IPA 1999) forms the basis for acoustic analysis.

Acoustic analysis includes the extraction of acoustic parameters of the speech signal using computational models. Features derived from the parameters can be used to create a statistical model of a speaker to account for the variability inherent to utterances of one speaker. These models are in turn compared against each other in a statistical evaluation[2]. An account of the acoustic properties used within acoustic analysis is given in section 2.2.2.

[2] It is important to consider the statistical distributions of the parameters at the beginning in order to decide on the applicability of the statistical models.

Rose (2002:35) states that '[...] the auditory analysis of a forensic sample is of equal importance to its acoustic analysis which the auditory analysis must logically precede'. The idea is that in order to proceed with a detailed acoustic analysis first a decision has to be made whether the recording can be used for identification at all, depending on its quality, and which parts or items of speech sounds can be compared against each other.

This highlights the need for using both kinds of analysis when dealing with forensic phonetic speech recordings. Jessen (2008) describes the use of both kinds of analysis for forensic speaker identification as follows.

> An acoustic-phonetic approach [. . .] builds upon an auditory-perceptual sound categorization and then investigates the acoustic manifestations of the perceptual categories. Acoustic phonetic analysis usually reveals that in acoustic reality, sound distinctions and sound separations in time are more gradient and less categorial in perception. Within a forensic context acoustic-phonetic analysis has the advantage that very accurate quantitative values can be provided, which would be impossible with auditory-perceptual analysis. However, it might not always be the case that additional accuracy actually increases the performance of speaker identification (Jessen 2008:17).

Automatic computerised systems constitute another kind of analysis form that relies entirely on statistical pattern recognition techniques applied to acoustic measurements. Speakers are statistically modelled using high-dimensional representations of features extracted from the speech recordings. These methods are used in biometric speaker verification and identification systems.

The statistical technique mostly applied are Gaussian Mixture Models (GMMs) where each feature vector dimension is modelled by a number of mixtures, i.e. sums of Gaussian distributions which represent the variation observed in the acoustical measurements. The measures used are described in detail in section 2.2.5.

Linguistic vs. non-linguistic parameters

Another distinction can be made between linguistic and non-linguistic parameters.

Features and cues in the speech sample can be linguistic in the sense of section 2.1.3 that they 'signal a contrast, either in the structure of a given language or across languages or dialects' (Rose 2002:44). An example for a linguistic auditory parameter is presented by a case where the realisations of a certain speech sound differ consistently between the samples being compared, which would imply a higher probability that the samples were spoken by different speakers than by the same.

Non-linguistic parameters are cues which are not relevant for the linguistic structure of the language being spoken. These include properties that signal the emotional state of the speaker, such as stress, fatigue etc.

2.2.1 Criteria for speech parameter selection

The choice of parameters in a particular case depends largely on the quantity and quality of the sound data available. For the development of new methods and to set criteria for testing them it is useful to consider which characteristics ideal speech parameter should attain.

Nolan (1983:11) proposes six characteristics that are highly desirable for forensic phonetic and acoustic parameters in general.

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

The first two criteria have already been discussed in section 2.1.3.

Resistance to attempted disguise or mimicry refers to the need for properties that are not easily manipulable by will. This is attained by parameters which are tied to the specifics of the speaker's physiology or typically go by 'unnoticed' in the attempt to imitate one's voice.

Availability refers to the need of parameters that can be gained from ordinary speech and do not rely on items or circumstances that are rather unlikely to appear in the samples used in a forensic case.

Robustness in transmission follows from the fact that the majority of forensic phonetic evidence descends from recordings of telephone speech which are limited in their frequency band. Parameters used for speaker discrimination should therefore remain unaffected by signal coding and recording to be of use, since comparability between speech sources must be maintained despite of different equipment (i.e. microphones) used. This affects especially modern telephone systems which are optimised to transmit spoken information in terms of parameters of a speech model, rather than speech as it can be recorded by a microphone.

The last characteristic, *measurability*, emphasises the need for parameters that can be extracted with relative ease. This is not restricted to manual extraction of features, which is very time-consuming, but also applies to automatic methods, e.g. if a parameter relies on finding the exact location of particular phonetic events, which cannot be done straightforwardly.

However, easy automatic measurement of features bears the risk of the features being used uncritically for estimating statistical models of speakers, despite not following the presupposed distribution functions.

In the following sections, a further distinction of acoustic features is being made which will segue from these general considerations into the characterisation of dynamic parameters that are used in the method outlined in chapter 3.

2.2.2 Traditional acoustic features

Rose (2002:41) defines traditional acoustic features as '[t]he acoustic cues that relate to differences between language sounds - either within a language or between languages'. These parameters have the beneficial property of being related in a straightforward way to the physiological basis of speech production: the different shapes and sizes of speakers' vocal tracts.

The features typically used in forensic phonetics are the fundamental frequency f_0 and the formant centre-frequencies F_i.

2.2.3 Fundamental frequency f_0

The fundamental frequency f_0 is described by Ladefoged (2000:164) as 'the number of complete repetitions (cycles) of variations in air pressure occurring in a second'. Consequently, it labels the frequency of opening and closing states of the glottis.

In the acoustic waveform of voiced sounds f_0 can be measured from the cycles in the quasi-periodic wave. It is often associated with pitch in the auditory domain which is correlated with the fundamental frequency.

It is considered a traditional acoustic parameter because it bears a linguistic function within language systems in that the presence of voicing indicates differences in meaning, e.g. between /s/ and /z/. The rate of vibration expresses linguistic contrast as well in that it signals stress. Furthermore, changes in fundamental frequency give rise to intonation.

The fundamental frequency f_0 has successfully been used for forensic speaker identification. The criteria for parameters (see section 2.2.1) are met in part, as it can be robustly extracted using auto-correlation techniques, and is readily available because voiced material is present in virtually every recording. However, the variability exhibited within and between speakers raises concerns as to its viability with regards to its use in forensic speaker recognition. Rose (2002:246) cites several factors influencing f_0 that were introduced in Braun (1995).

- *Technical factors* (sample size, tape speed)
- *Physiological factors* (race, age, smoking and intoxication)
- *Psychological factors* (emotional state and situational factors, including background noise level and time of day)

This summary, however, delivers a very imprecise picture of factors and is at least debatable if not precarious. First of all the claim of race as a physiological factor cannot be upheld on a scientific basis. Whilst there are studies[3] of vocal tract dimension which claim that, beside gender, race is ‘one of the most important factors affecting the oral and nasal structures’ (Xue & Hao 2006:392), this cannot readily be relayed to factors of variability in f_0.

Furthermore, situational factors include aspects that cannot be subsumed under psychological factors and, thus, deserves a category of its own. Individual language differences present another category along which the fundamental frequency varies, as it fulfils various differing functions within the language’s system.

Hence, to control for the factors that generate variability in the domain of f_0, sociological factors such as gender, situation, and sociolects must be determined. These aspects must be controlled for in order to ensure a legitimate basis

[3] The study measures vocal tract dimension by acoustic pharyngometry. The subjects were controlled for age, gender, height and weight. ‘Race was found to be a significant variable for oral volume and total vocal tract volume’ (Xue & Hao 2006:395)

for forensic speaker comparison and emphasise the importance of preliminary auditory analysis.

2.2.4 Formants F_i

The formants F_i represent the acoustic resonances produced by the dynamics of the vocal tract. As already mentioned, the formant values exhibit correlation with the production and perception of speech sounds. The formant centre-frequencies are usually given by the maximum amplitudes in the LPC spectrum of a speech sound which results from specific articulatory vocal tract settings.

The aforementioned *source-filter theory of speech production* (Fant 1960) as well as the *perturbation theory* (Chiba & Kajiyama 1958) present approaches for relating formant frequencies to the articulatory state of the vocal tract. These two theories are described below.

The tube model

The vocal tract is modelled as a series of uniform cross-sectional tubes. The formant frequencies can be calculated and therefore predicted given the length of each tube using the formula $F_n = \frac{(2n-1) \cdot c}{4 \cdot length}$, where c is the speed of sound.

Figure 2.2 provides an illustration of a tube model for the vocal tract configuration for [ɑ].

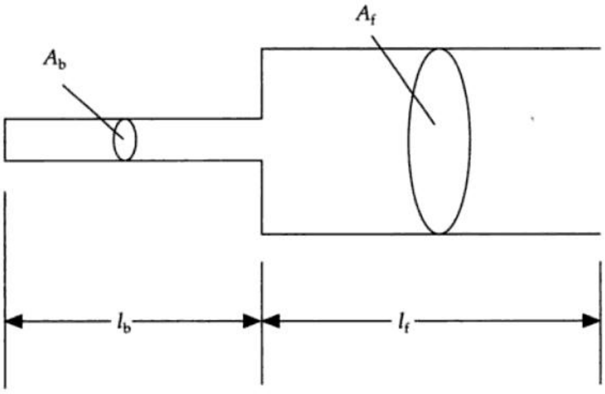


Figure 2.2: A two tube model approximation of the vocal tract for [ɑ] (Johnson 2003)

The model gives support for the assumption that formant centre-frequencies of a speech sound are related to the characteristics of the speaker's vocal tract in that it relates the physical dimensions of the idealised tubes between the larynx and the lips to the acoustic output.

It must be noted, however, that the example given in figure 2.2 is the most basic configuration and that sounds exist that cannot be modelled by this approach due to its inherent limitations[4].

The perturbation theory

The perturbation theory models vowel acoustics using the relationship between air pressure and velocity. The consequences of vocal tract constrictions on formant frequencies are summarized in (Johnson 2003:110).

> The perturbation theory [. . .] relates vocal tract constrictions to formant frequencies by taking into account the kinetic energy present at points of maximum velocity and the potential energy present at points of maximum pressure

Figure 2.3 shows the locations of the points of maximum velocity (N_i) and maximum pressure (R_i) in a straight tube and the relation to the human vocal tract.

Both theories describe the correlation between the positioning of articulators and hence the properties of the speaker's vocal tract, and the formant frequencies in the acoustic domain.

The relationship between formant frequencies and articulation

Since the inception of acoustic phonetic analysis techniques a pursuit was undertaken to find a model that relates acoustic cues with vocal tract settings and articulation. Correlation between the first and second formant and the tongue tip position was noted early on. 'The convention of representing the formant data of vowels in an F1/F2 plot goes back to Joos (1948) [. . .]' (Moosmüller 2007b:31). This relationship was presented in a plot of frequencies of the first and second formant of different vowels in which the 'the scales [. . .] were deliberately set up so as to enhance the resemblance of the acoustic chart to the tongue-position chart' (Joos 1948:53).

[4] See for example Holmes (2001) for a discussion of problems related to the modelling of higher resonances.

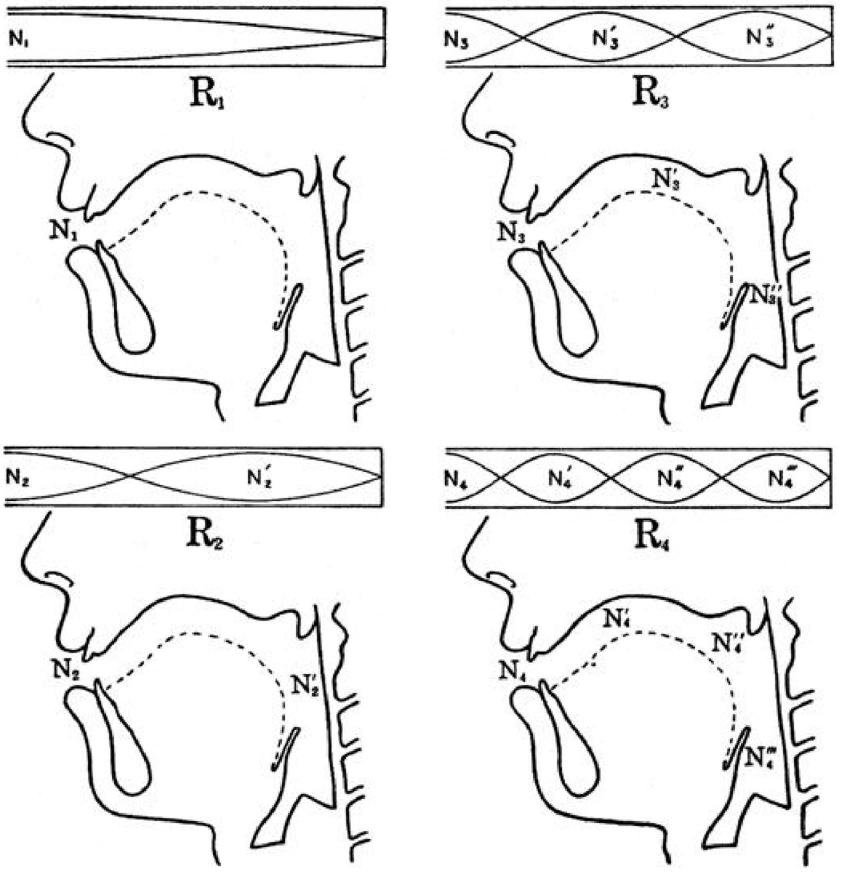
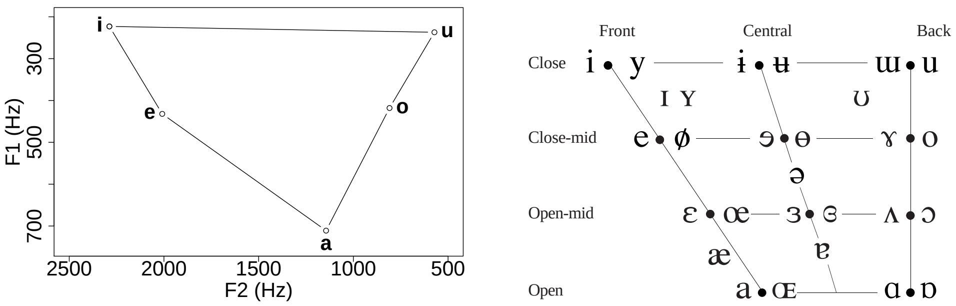


Figure 2.3: A depiction of the points of maximum velocity (N_i') and pressure (R_i) in perturbation theory (after Chiba & Kajiyama (1958))

Figure 2.2.4 shows the simplified relation between the formant frequencies and the articulation as it is still commonly described. This model is useful to somewhat characterise tendencies in formant behaviour with respect to the position of the tongue, yet fails to give an accurate depiction of the aspects that have an effect on formant frequencies, such as lip rounding and protrusion Ladefoged (2000:35).

As already brought forward by Fant (1960:11) '[t]he highest point of the tongue is well correlated with the relevant acoustic data but does not uniquely define the resonator dimensions'. Thus, the IPA quadrilateral should not be seen as being strictly based on articulation, but also on auditory and acoustical definitions (IPA 1999:12).

Further evidence against this simplistic picture is brought forward by the



(a) Acoustic vowel space

(b) The IPA vowel chart

Figure 2.4: Simplified relation between formant frequencies & articulation

quantal theory of speech production (Stevens 1989) which suggests a non-linear relation between the acoustic and the auditory domain by defining three zones of acoustical stability under differing articulation.

These and many other aspects[5] show that the purported relationship established using the acoustic vowel space and the vowel quadrilateral cannot be upheld on a scientific basis. Nevertheless, there exists of course a relationship between articulation and acoustics which is of a more complex nature.

This relationship has its merits with regards to the use of these parameters in court. As Alderman (2005:13) notes '[t]he correlations between formants and physiology are supposed to make the concepts more understandable to laypersons, such as members of the jury, or even magistrates, judges and lawyers, and thus make the deciphering of expert evidence an easier task'.

In practice, formant frequencies are extracted from speech recordings by formant tracking algorithms. As formant frequencies build the basis of the parameters used in the present work one common algorithm is briefly outlined in chapter 3.

Limitations of traditional acoustic parameters

In forensic practice, a severe limitation exists when using formant frequencies, as the overwhelming amount of speech samples under considerations are recordings of telephone conversations. The technical constraints involved consequently limit the bandwidth of frequencies transmitted over telephone networks to approximately 300-3400 Hz, which renders the formants above the third and usually

[5] For a thorough discussion see Moosmüller (2007b:32).

also the first formant virtually useless for speaker comparison. This additionally decreases the dimensions which speakers can be discriminated in, as higher-frequency formants are regarded as bearing more speaker-specific information because they '[...] often reflect the resonances of relatively fixed smaller cavities in the vocal tract, for example the larynx tube, which are assumed to be relatively unaffected by the gross configurational changes of the vocal tract [...]' (Rose 2002:237).

Furthermore, several studies suggest the existence of a so called *telephone effect* (Künzel 2001), in which formant frequencies are shifted in a non-uniform way, rendering speaker comparison based on formants a delicate task. The causes and extent of this phenomenon are still rather unclear. Fecher (2008:82), while studying the effects of Voice-over-IP transmission technologies on traditional acoustic-phonetic parameters, noted band-pass filtering of the signal as a possible cause.

Guillemin & Watson (2008) examined the effects of coders used in the GSM mobile phone network on the speech signal. They applied each coder for the whole speech sample and found significant impact on formant frequencies, especially for low pitch male speech. The situation is exacerbated by the fact that, in real GSM telephone transmissions, the coder can be changed every 20 ms to compensate for poor channel conditions (Guillemin & Watson 2008:300) Furthermore, if packet loss occurs, mechanisms are employed that interpolate the signal or reinsert the last speech frame, leading to a speech signal that partially differs from the original.

The range of traditional features can be subdivided based on their role within the linguistic system of the language at hand.

Acoustic linguistic parameters

Acoustic linguistic parameters subsume the use of traditional features like the fundamental frequency and formant centre-frequencies in forensic speaker recognition, taking into account their role in the linguistic system.

As described in the previous sections, formant frequency values result from the physiological condition and articulatory setting of the speech organs during the production of speech sounds. Speakers' formant frequencies differ because of their physiology and habitual aspects in the movement of the articulators while achieving phonetic targets, but they exhibit some degree of within-speaker variability.

This variability and the behaviour of formant frequencies in general depend on the linguistic item to be produced and of course on its context. To compare recordings of speakers, comparable units have to be found to perform an acoustic linguistic analysis. These must be controlled for similar context and stress position to limit the variability exhibited by the parameters within each speaker.

Formants have long been used in this regard. Acoustic analysis based on traditional acoustic parameters derived from a vowel is often conducted by either calculating the mean of the formant measurements or by using a measurement during the steady state (usually near or at the midpoint) of the vowel for comparison. The latter is based on the notion of phonetic targets.

Extending the account given in section 2.1.3, the notion of the phonetic target is more thoroughly discussed. A definition can be found in Lindblom (1963).

> A target is specified by the asymptotic values of the first two formant frequencies of the vowel [...] (Lindblom 1963:1773).
>
> A target was found to be independent of consonantal context and duration and can thus be looked upon as an invariant attribute of the vowel. Although a phoneme can be realized in a more or less reduced fashion, the talker's "intention" that underlies the pronunciation of the vowel is always the same, independent of contextual circumstances. A vowel target appears to represent some physiological invariance (Lindblom 1963:1778).

Moosmüller (2007b:174) argues that this definition essentially implies that 'the target is identical with the phoneme' and, thus, 'more or less to a pronunciation under ideal conditions'. This, however, leads to problems regarding the variability of speech which was solved by introducing the concept of target undershoot in an attempt to account for the fact that the allegedly contextual invariant target is almost never reached.

Other characterisations and solutions have been put forward, including *the window model of coarticulation* which proposes an articulatory window for features, a 'range of minimum and maximum value that the observed values must fall within' (Keating 1990:455).

Moosmüller (2007a) delineates the approach within the framework of *Natural Phonology* (Donegan & Stampe 1979, Dressler 1984).

> In Natural Phonology, the phoneme is an invariant mental representation of a sound, and the way from phoneme to phonetic output is

> determined by phonological processes, which are, in any case, phonetically motivated and which follow certain universal preferencies, e.g. the preference for figure-ground contour sharpening (Moosmüller 2007a:498).

If and to what extent these phonological processes are applied is, however, language dependent. Consequently, the target is variant whereas the phoneme is invariant.

The notion adopted in the present work follows this view of phonetic targets.

> [. . .] there is no standard method for identifying where the vowel target occurs partly because many monophthongal vowels often have no clearly identifiable steady-state or else the steady-state, or interval that changes the least, may be different for different formants. (Harrington in press:85)

These measurements are in turn used for comparing segments of different speakers.

Acoustic non-linguistic parameters

Acoustic non-linguistic parameters can be characterised as features that have no inherent function within a single language system. These are usually based on averaging over the traditional acoustic parameter measurements of whole utterances.

The long-term average f_0 (*LTF0*) method is used for the comparison of speakers based on the statistical distributions of their fundamental frequency values. As noted in Rose (2006b), it depends on and reflects non-linguistic information which, using Nolan's voice model (see section 2.1.4), can be ascribed to the indexical factors like the state of health and physiological aspects as well as affect and self-presentation. However, studies by Kinoshita (2005) show that the use of the LTF0 method is rather limited due to the wide range of variation of the f_0 parameter within a speaker, especially if the recording was taken under detrimental circumstances like noise which causes increased vocal effort and thus a rise in mean f_0 (Jessen et al. 2005). Mismatch of speaking style among the comparison samples also has a strong deteriorating effect on the performance of these methods (Becker 2008).

The long-term formant distribution (*LTF*) is a method that averages over the formant values of all vowels produced by a speaker. For each formant in a recording a LTF value is calculated by taking the arithmetic mean of all formant centre-frequency measurements. Studies based on this method (Nolan & Grigoras 2005, Grigoras 2006) show that LTF satisfies the criteria proposed by (Nolan 1983) (see section 2.2). A study by Moos (2008) investigated the applicability of this method on samples taken under different recording and transmission conditions and concluded that separation of the speakers was attainable. However, as with most other (semi-)automatic or summary methods, this approach has to be used in combination with other procedures in order to gain a more complete picture of speaker differences. Grigoras et al. (2009) compares two kinds of LTF methods with other automatic approaches and found their performance close to GMM-based methods (see section 2.2.5).

Another acoustic non-linguistic method for forensic speaker recognition is called the long-term (average) spectrum (LTS or LTAS) (Nolan 1983:130). It is calculated by taking the average of a series of short time spectra, which results in a measure of the distribution of acoustic energy. Studies employing this method have shown that several criteria for forensic phonetic parameters are not met. Rose (2002:261) notes its sensitivity to voice disguise and channel mismatch as well as substantial inter-session variability if the samples originate from sessions several days apart.

Lindh (2004), however, ascribes ‘promising performance’ to a method based on graphic representations of LTAS. The study involved closed-set speaker identification tests using recordings of speakers employing different kinds of voice disguise like dialect, accent, whisper and falsetto.

2.2.5 Automatic-acoustic features

Automatic-acoustic features are parameters extracted automatically from the signal by a computer algorithm. They form the basis of commercial biometric automatic speaker verification and identification systems used for physical access control or voice authentication over telephone. Only recently they have started being used in the forensic domain.

The most commonly used parameters are derived from the *cepstrum* of a signal (Bogert et al. 1963). Its initial use was the estimation of f_0 because of the fact that it 'effectively decoupled the part of the speech wave that were due to the glottal excitation from those that were due to the supralaryngeal response' (Rose 2002:262). Later it was applied to speaker as well as speech recognition.

The most widely used automatic-acoustic parameters are Mel Frequency Cepstrum Coefficients (MFCCs), which are similar to the Cepstrum, but the frequency scale used for the calculation of the MFCCs is not in Hertz (Hz) but in *Mel* which is a perceptual unit of pitch (Stevens et al. 1937).

They are derived from a signal by applying the following steps.

1. *Amplitude normalisation*
   To compensate for absolute acoustic energy differences the average amplitude of all samples in the signal is subtracted from each sample and subsequently divided by the maximum amplitude.
2. *Windowing*
   The signal is divided into frames of equal length, often overlapping each other by a specified amount. Subsequently a windowing procedure like the *Hanning* window is applied to the samples of each frame.
3. *Fourier transform*
   The Fourier transform is applied to each frame to calculate the spectrum which is in turn squared to arrive at the power spectrum.
4. *Mel frequency scale*
   The power spectrum is then warped to the Mel frequency scale and logarithmised.
5. *Discrete Cosine Transform*
   The discrete cosine transform is applied to the resulting logarithmised and Mel-transformed spectrum.

The amplitudes of the resulting spectrum constitute the MFCCs. The effect of applying this method to the signal is to smooth its spectrum which leads to the discount of frequency components that are introduced by noise. This quite abstract representation of the speech signal has been shown to be very successfully applicable to speaker identification tasks, resulting in very low error rates.

One critical point is that there does not seem to exist a specific relation of the particular MFCC coefficients to the perceptual properties of the speech apparatus of the respective speaker. It has been shown that certain coefficients correlate to aspects of the vocal tract (Rose 2002), however the interpretation of these values is not straightforward. However, an explanation of why the combination of a perceptual scale (Mel) with the smoothing of amplitude spectra performs better than alternative scales to be applied is still lacking.

2.2.6 Dynamic features based on traditional acoustic features

The central notion assumed in this study is that speakers are less constrained in their articulatory movements and behaviour while they move from one phonetic target to the next. They carry out a phonetic plan which, according to Nolan's model of the information comprising a voice, is the outcome of the linguistic mechanism and the speaker's communicative intent.

Within the framework of Natural Phonology the subject area is contrived as part of social interaction. Hence, '[t]he two main functions of segmental phonology are to make language pronounceable and perceptible' (Dressler 1984:32).

Moosmüller (1997b:32) elaborates the relationship between phonemes as phonological units and the speaker's intention.

> Following Baudouin de Courtenay (1894), phonemes, the outputs of language-specific processes (based on universal phonological processes), are defined as intentions. Any intended phoneme is accompanied by an additional social intention. The phonetic output may diverge from this intention, phonologically (in the sense of a phonological process), socially (in the sense of a variety not intended) or both (in the sense of socially evaluated processes) (Moosmüller 1997b:32).

Following the train of thought that speakers can be identified by the dynamics exhibited during the realisation of their intention, this framework provides useful, as it incorporates these aspects to account for factors of variability.

Regarding the phonetic implementation within Natural Phonology, Donegan (2002:58) states that '[t]he phonological representations specify combinations of features in relative time, rather like a musical score, and the vocal organs 'interpolate' as they move from one target or gesture to the next'. However, '[s]peakers do not simply line up a sequence of phonemic targets and allow the articulators to get from one to another as best as they can; instead the activity of articulation is centrally planned, so that features spread (or gestures overlap) in regular ways' (Donegan 2002:69).

This is supported by the findings of a study by Whalen (1990) which investigated coarticulation effects by requiring test persons to start reading nonsense strings in which consonants and vowels were inserted only after the speaker began to read, i.e. before the whole utterance was shown. The author concluded that '[c]oarticulation, though presumed to be due to the constraints of producing speech in real time, is largely a result of planning an utterance rather than an automatic consequence of successfully producing that utterance' (Whalen 1990:29).

As has been elaborated in the preceding sections there exist differences between humans with respect to their physique that has effects on the dynamic properties of speech. Nolan (1983:60-61) states that '[...] it is reasonable to assume that different speakers may have differential agility in speech production, in the same way that speed of movement and coordination differ in other physical skills [...]'.

In the pursuit of finding useful parameters for robust forensic speaker recognition based on speech segments, the notion of phonetic targets (see section 2.2.4) has been adopted. In the case of vowels these are usually stated in terms of formant frequencies that have to be maintained at some point in time during the segment in order to enable the perception of the phonological unit. In monophthongal vowels only one phonetic target is assumed that 'can be thought of as a single point that [...] typically occurs nearest near the temporal midpoint [...]' (Harrington in press:85). The onset and offset of the vowel are subject to coarticulation effects depending on the context. Thus the time-dynamic properties of the formant features under consideration are to a large extent subject to the surroundings of the vowel.

In the case of diphthongs the assumption was made that two targets were involved in the production which have to be achieved to attain correct perception. A study by Watson & Harrington (1999) showed, however, that these targets are not sufficient on their own to allow discrimination between diphthongs and vowels in general, indicating that the linguistic information is conveyed by means of other more dynamic properties, as 'vowels can vary in length, in the relative timing of the target, and in whether vowels are specified by one target or two' (Harrington in press:88).

Diphthongs have often been characterised by and divided into an onset steady state, the glide and an offset steady state, assuming that relative timing and duration are decisive factors for the language-specific diphthong perception and discrimination. However, when applied to real formant data, ‘steady state’ is often a rather inappropriate term. For example McDougall (2005:51) notes that her Australian English /aɪ/ data rather shows ‘a relatively steady onset component followed by a strong glide movement’, which applies to Austrian Standard German /aɛ/ data as well.

The fact that diphthongs occur with a relatively high frequency in speech and that the dynamic properties exhibited in the spectral change over time are measurable in a straightforward way are properties that place them in a position of high interest for forensic speaker recognition.

Several studies have investigated the use of diphthongs for discriminating speakers. Previous research concentrated mostly on instantaneous features to capture the notion of phonetic targets and their realisation. In monophthongs or liquids this approach was quite successfully applied by concentrating on parameters taken from the steady state or by calculating the mean of the measured values (see Rose (2006a), Rose et al. (2006), Alderman (2005)).

However, as noted by Kinoshita & Osanai (2006:112), the formant contours exhibit substantial style-specific behaviour and are, thus, subject to rather high inter-session variability which renders methods depending on formant values of the two targets inapplicable. Kinoshita & Osanai correspondingly investigate the use of other features derived from the formant trajectories. They use a likelihood ratio approach (see section 2.3.2) to evaluate combining formant target values with the slope of the glide of the second formant, yet conclude that ‘the slope of F2 was not found to be particularly robust against differences in speech styles. However, the angle of the glide was at least as useful as the two targets of the diphthongs [. . .]’ (Kinoshita & Osanai 2006:117).

Recent studies employed more complex parametric representations of formant trajectories based on parametric functions fitted to formant contours. McDougall (2005) first described the use of linear regression techniques to adapt polynomial functions of different degrees to vowel trajectories in order to discriminate between speakers. In her study she investigated several methods to characterise dynamic properties of speech based on formant frequency measurements of /aɪk/

and of the coarticulation between vowels and schwa as well as intervocalic /r/ produced by speakers of Standard Southern British English (SSBE). The methods used were measurements at temporal midpoints, measurements taken at 10 percent steps throughout the diphthong, as well as linear regression to fit polynomials to formant trajectories.

McDougall & Nolan (2007) extended this approach by fitting polynomial functions to formant measurements at 10% intervals obtained from /uː/ produced by male speakers of Standard Southern British English (SSBE). They performed discriminant analysis to find the best-performing parametric representation. The results indicated that the quadratic polynomial best captures speaker-specific dynamic properties.

Studies conducted by Morrison & Kinoshita (2008), Morrison (2008), and Morrison (2009b) made use of parametric representations along the lines of McDougall & Nolan (2007) combined with a likelihood ratio approach (see section 2.3.2) to perform speaker discrimination tests based on data of male speakers of Australian English.

Morrison (2008) investigated the use of parametric representations for forensic speaker comparison based on recordings of 27 male speakers of Australian English. The samples were taken in two sessions where the speakers were asked to read sentences like ‘bide, B-I-D-E spells bide’. Of each sentence, two recordings were made in each session. He then used quadratic and cubic polynomial functions to model the formant dynamics exhibited during the production of /aɪ/ and compared the obtained likelihood ratio scores with those obtained by applying traditional dual-target approaches. The results showed that the parametric representations outperformed other methods.

Morrison & Kinoshita (2008) again used audio recordings collected from 27 male speakers of Australian English aged 20 to 63 years who were asked to speak sentences of the form ”Hoe, H-O-E spells hoe.” The /oʊ/ diphthong of the first and final word were segmented manually. The formant frequency trajectories of the first three formants were tracked using a standard formant tracking algorithm and were manually corrected where necessary.

Quadratic and cubic polynomial functions as well as the first three and four coefficients derived from discrete cosine transform (DCT) were used as parametric representations of the diphthongal formant contours. The two kinds of parametric representations were compared against each other in terms of the C_{llr} metric (see section 3.4.6) with respect to several conditions, namely applying time-normalisation to the formant contours as well as using a logarithmic frequency scale. The likelihood ratios were calibrated post-hoc by applying linear calibration techniques (see section 3.5).

Morrison (2009b) finally also used recordings of 27 male Australian English speakers who read similar sentences containing the diphthongs /aɪ/, /eɪ/, /oʊ/, /aʊ/, and /ɔɪ/. He used the same types of representations as in the previous study. The resulting likelihood ratios from the individual segments were once again calibrated as well as fused using logistic regression fusion (see section 3.6).

2.3 Evaluation of forensic speech evidence

Acknowledging the problems posed by forensic speaker recognition, amongst others by between- and within-speaker variability, the question arises as to how forensic speech evidence should be evaluated and interpreted in court.

As noted by Aitken (1995:4), scientific observations give rise to random variation. The resulting uncertainty must be accounted for by using probabilistic and statistical measures when assessing the strength of evidence.

In criminalistics, however, the concept of identification used in court or by the police prefers individualisation, i.e. a categorial decision of similarity or dissimilarity, guilt or innocence. Strictly speaking the certainty of the decision which is strived for requires a feature that is so rare that it can be concluded that there exists only one person bearing that feature, as it is the case with fingerprint analysis (Champod & Meuwly 2000:1).

Nolan (2001) states that it has yet not been scientifically proven ‘whether absolute discrimination is even theoretically attainable’, as speakers’ ranges of variation overlap in the multidimensional feature space. However, as noted in section 2.1.2, phoneticians nevertheless should offer their opinion before court, but it comes down to how to express their testimony.

The following section discusses an approach which tries to solve the situation. It is commonly known as the Bayesian approach and represents the framework which is adopted in this thesis.

2.3.1 The Bayesian approach

The Bayesian approach provides a conceptual framework of how to evaluate the strength of evidence given two competing hypotheses. The name is taken from the Bayes' Theorem (Bayes 1763) which basically allows to inverse conditional probabilities.

The question usually asked in court is most commonly phrased as 'Is the speaker heard on the incriminating recording the defendant?', or quite similar 'How probable is it that the offending sample comes from the defendant?'. As noted in Rose (2002:56) this question cannot be answered by the forensic phonetic expert for logical and legal reasons.

First of all the probability of the offender being the defendant cannot be stated for *legal reasons*, as the forensic phonetician exceeds his authority and role in the judicial process, as it is the role of the judge or jury to reach a decision of guilt or innocence. Rather he should be concerned with giving an assessment of the strength of the evidence (Rose 2006a:64).

Then, on the basis of *logical reasons* he cannot make a statement of probability concerning the identity of speakers. The forensic expert does not have access to all the information available to the judge or jury that is necessary to make that statement, as there could be strong evidence otherwise against the defendant's involvement in the crime (Robertson & Vignaux 1995).

The Bayesian approach alleviates the situation by making explicit the role of the forensic scientist and to allow for easy combination of different types of forensic evidence. First it is acknowledged and made explicit that there exist two hypotheses concerning the guilt or innocence of the defendant.

- The prosecution hypothesis, denoted as H_{SS}, represents the claim that the speech on the offending recording originates from the defendant.
- The hypothesis of the defence, denoted as H_{DS}, states that there are different speakers involved.

The question asked by the court is now rephrased as a ratio of probability of the two competing hypotheses given the evidence of the forensic scientist.

$$\begin{array}{ccccc} \frac{p(H_{SS}|E_{Sp})}{p(H_{DS}|E_{Sp})} & = & \frac{p(H_{SS})}{p(H_{DS})} & \cdot & \frac{p(E_{Sp}|H_{SS})}{p(E_{Sp}|H_{DS})} \\ \text{Posterior Odds} & & \text{Prior Odds} & & \text{Likelihood Ratio} \end{array} \tag{2.1}$$

Equation 2.1 shows the Bayes' Theorem in its odds form, as it is applied in the case of forensic speech evidence (Rose 2002:63). E_{Sp} denotes the speech evidence while H_{SS} and H_{DS} represent the prosecution and defence hypotheses, respectively.

Odds are basically the same as probability but expressed in a slightly different but often more comprehensible form. For example, an event that occurs with the probability of 75% is expressed in odds as 3:1, which means that it is three times more likely to happen than not to happen. The expression $\frac{p(E)}{1-p(E)}$ performs a conversion between probabilities and odds.

Central to the Bayes' Theorem is that the *prior odds*, which are the odds in favour of the prosecution hypothesis against the defence hypothesis before the evidence is considered, are updated to *posterior odds* by multiplying them by the *likelihood ratio*, which is 'the ratio of the probabilities of evidence assuming guilt and assuming innocence of the suspect' (Aitken 1995:46). The role of the forensic phonetic expert is to provide his assessment of the strength of evidence expressed as a likelihood ratio value.

As can be seen from the formula it is not sufficient to only look at the probability of evidence under the prosecution hypothesis or only under the defence hypothesis. The following section explains this reasoning and further describes the concept of the likelihood ratio.

2.3.2 The likelihood ratio

As introduced in the previous section, the concept of the likelihood ratio is the practical solution to the question of how to make a logically and legally correct assessment of the speech evidence by the forensic expert, as it provides a continuous numerical expression of the strength of evidence under consideration.

In the calculation of the likelihood ratio the forensic practitioner expresses the ratio of the probability of evidence assuming the prosecution hypothesis H_{SS}, that is, the samples originate from the same speaker, and the probability of evidence given the hypothesis of the defence, usually that the samples originate from different speakers (Aitken 1995, Lindley 1977). However, it can also be specified in a different way, as explained below.

Equation 2.2 shows how the likelihood ratio is expressed.

$$LR = \frac{p(E_{Sp}|H_{SS})}{p(E_{Sp}|H_{DS})} \tag{2.2}$$

E_{Sp} again denotes the speech evidence while H_{SS} and H_{DS} represent the prosecution and defence hypotheses, respectively.

The expression can be seen as a balance of *similarity* to *typicality*[6] (Rose 2006b:168). In the numerator, a score is given of how similar the parameters of the evidence are, and in the denominator a value is derived for how likely it is to find the evidence in a specified *reference population.* This in turn depends on the hypothesis of the defence, which can read simply as 'it was a different speaker', or be more specific as in 'it was the accused's brother'. This concept of balance of probabilities is vital to the likelihood ratio as there are properties which distinguish speakers more certain than others.

A problem remains in the process of defining a reference population. If the defence states that it was the accused's brother then it is very simple, as the objective turns into the identification of a speaker in a closed set. Yet, if the assertion is made that it was a different speaker who sounds similar to the accused then the task becomes less trivial, as it is difficult to obtain data from speakers that fit that criterion for each case. A possible approximation for this hypothesis is to use the data of speakers of similar sex, age and body height with the same first language and possibly dialect as a reference population which relate to the acoustic parameters discussed in section 2.2.2.

However, as the judge or jury is not supposed to be given evidence in form of numerical representations like the likelihood ratio scales of verbal equivalents have been proposed that re-introduce a rather categorial notion but are in turn more readily understandable by the finder-of-fact. Table 2.2 shows such a list of verbal equivalents used at the Forensic Science Service as it is presented in Rose (2002:61).

[6] It must be noted that both terms do not equally correspond to homophonous concepts in statistical nomenclature and should not be taken as denoting the same strict formalisation as in their use in statistics. They should rather imply a more intuitive understanding of the purpose of the likelihood ratio concept.

LR value	Verbal equivalent
> 10000	Very strong evidence
1000-10000	Strong evidence
100-1000	Moderately strong evidence
10-100	Moderate evidence
1-10	Limited evidence
	supporting same speaker hypothesis
1-0.1	Limited evidence
0.1-0.01	Moderate evidence
0.01-0.001	Moderately strong evidence
0.001-0.0001	Strong evidence
< 0.0001	Very strong evidence
	against same speaker hypothesis

Table 2.2: Verbal equivalents of likelihood ratio values (Rose 2002:61)

Logarithm of the likelihood ratio

Taking the logarithm of the likelihood ratio has theoretical as well as practical advantages.

The theoretical gain from using a logarithmic form is related with the interpretation of the value as a measure of strength of evidence. As noted in Aitken (1995:45) the prior probability in favour of H_{SS} are multiplied with the likelihood ratio in the odds form of the Bayes' theorem (see equation 2.1). By applying the logarithm this turns into an additive relationship which facilitates the interpretation of the likelihood ratio as a weight. As Aitken puts it, '[a] positive weight may be thought to tip the scales of justice one way, a negative weight may be thought to tip the scales of justice the other way'. A likelihood ratio of one becomes zero when the logarithm is applied, leading to the correct interpretation of neither adding weight to the prosecution hypothesis nor to the hypothesis of the defence.

The practical use is to increase numerical precision in the calculation of the likelihood ratio using computers, as the resulting values can theoretically get infinitely large or small. Taking the logarithm alleviates this problem as the resulting values are scaled down. Furthermore, uncorrelated likelihood ratios from several methods can be combined by multiplication, which can lead to loss of precision due to underflow in the floating point presentation used in computers if very low likelihood ratios are involved.

2.3.3 Discussion

The explicitness of the Bayesian approach prevents a wide range of errors in interpretation, including the following fallacies noted in (Aitken 1995:36-44).

- *Fallacy of the transposed conditional*
  This fallacy describes the case where the probability of evidence assuming the prosecution hypothesis $p(E_{Sp}|H_{SS})$ is calculated but is taken to be the probability of the prosecution hypothesis given the evidence $p(H_{SS}|E_{Sp})$.

- *Defender's fallacy*
  This type of error occurs if it is stated that the evidence has little relevance because the suspect is one of a rather large number of people with a similar property, but it is neglected that before the adduction of evidence the prior probability would have been nearly nil.

- *Probability (another match) error*
  In this common fallacy the probability of evidence assuming the defence hypothesis is equated to the probability that at least one other person has the same property.

As the consequences following from these misinterpretations can be rather grave, the benefits of stating the evidence as a likelihood ratio instead of some other form of probability statement become clear.

The application of the Bayesian approach to the problem of evaluating evidence has not yet been embraced by the whole community of forensic phonetic theorists and practitioners. Currently its utilisation is strongly debated by proponents and opponents working in the field.

The UK Position Statement on forensic speaker comparison accepts the presupposition of the Bayesian approach that it is not the forensic expert's role to make an identification claim, but rather to perform a speaker comparison (French & Harrison 2007:138), however finally rejects the framework as it was presented here because of the lack of data for reference populations and proposes a two-stage speaker comparison procedure. Rose & Morrison (2009), in a response to the UK Position Statement, criticise this proposal in that it in fact faces the same problem with respect to reference populations and the distribution of forensic phonetic parameters within the demographic.

A central difficulty stated by many opponents of this approach is how to assess the prior odds in practice. This problem is well-recognised, as Rose (2002:73-74) discusses this question but finally does not state a solution. However, the role of the forensic phonetician within the Bayesian framework is to calculate the likelihood ratio, which can nonetheless be reported in court as the strength of evidence.

The very basis of the Bayesian approach, that is the combination of prior odds with an estimate of the strength of evidence given by the likelihood ratio is incompatible with many legislations around the world, especially in western Europe. This is because of the fact that the finder-of-fact must specify prior odds which basically state how likely it is that the suspect is actually the offender, prior to adducting the information gained by the evidence. This, of course, collides with the presumption of innocence. But, as argued by Rose (2002:75) it can be 'shown by NSM[7] analysis that incompatibility of prior odds with presumption of innocence is not a valid criticism of the legal use of Bayes' theorem'.

Given the criticism presented above, a comparison with other probability statements for speaker identification is due. Statistical tests of significance presuppose a null hypothesis which usually states that there is no difference in the mean and variance of parameters of both speakers. Aitken & Lucy (2004:112) cites a method involving multiple significance tests where each parameter is individually tested. The prosecution (null) hypothesis is rejected 'if any of the individual variable mean differences is greater than three standard errors', under the assumption that the parameters are uncorrelated, i.e. statistically independent. An other test presented is Hotelling's T^2-test, a multivariate generalisation of Student's t-test.

These kinds of tests share the problem concerning the presumption of innocence, as the null hypothesis states that both speakers are the same. Furthermore, the principal difference between likelihood ratio tests and significance tests is that in the latter, the prosecution hypothesis can only be rejected but, in a strict sense, not verified, whereas the former signals the strength of support for each hypothesis by it's deviation from unity (or zero, for log likelihood ratios). That is because statistical tests never introduce causality, but provide a decision between two hypotheses.

[7] Natural Semantic Metalanguage

However, the deviation of likelihood ratios from one is also the source of another point of criticism, as it necessarily does not have an upper and lower bound. Therefore it is difficult to interpret the scores and one has to adhere to the range of scores given during extensive tests with very similar cases and conditions to rely on its validity.

The straightforward combination of different sources of evidence through multiplication of likelihood ratios is cited by Robertson & Vignaux (1997) as another practical advantage of this approach, who state that '[s]ignificance tests and probabilities of paternity cannot logically be combined with other evidence at all'.

Nevertheless, many forensic practitioners see a coming paradigm shift, not only regarding the evaluation and presentation of speech evidence but of forensic evidence in general to build a framework for forensic identification based on rigorously tested methods. Saks & Koehler (2005:892) note that forensic science was based on the assumption that 'two indistinguishable marks must have been produced by a single object [. . .], leaning on the assumption of discernible uniqueness'. Yet, because of experience gained by DNA typing and a change in legal admissibility standards, the forensic sciences are undergoing a paradigm shift.

During the rise of DNA analysis this new technique was applied to cases where a suspect had already been convicted. Further inspection of 86 cases where the DNA tests resulted in post-conviction exonerations showed that 'forensic science expert testimony is the second most common contributing factor to wrongful convictions, found in 63% of those cases' (Saks & Koehler 2005:893).

Thus, forensic practitioners including scientists from the phonetic domain are pushing towards the adoption of a more rigorous approach to evaluating evidence based on population statistics (Drygajlo 2007, Morrison 2009a). Gonzalez-Rodriguez et al. (2007) present the results of different forensic acoustic-phonetic systems and analyses based on the Bayesian approach in order to emulate DNA-like transparency and testability of the methods. They calculate likelihood ratios using traditional forensic phonetic features using a generative likelihood ratio formula developed by Aitken & Lucy (2004) (see section 3.3) as well as an automatic forensic speaker recognition system with scores calibrated to give likelihood ratio outputs.

Eckert & Wright (1997:79) state the following test as to 'whether the science or scientific tests employed are of such a level of validity as to be allowed into evidence':

1. Whether the type of evidence can be and has been tested by scientific methodology
2. Whether the underlying theory or techniques has been subjected to peer review and has been published in the professional literature (although this is not a sine qua non)
3. How reliable the results are in terms of potential error rate
4. General acceptance (the old Frye test) can have a bearing on the inquiry

Consequently, the aim of forensic speaker recognition must be to attain the level given by this test. The calculation of a likelihood ratio in terms of a Bayesian approach accomplishes this by being easily testable, yet has to be verified based on other languages and background populations in order to ascertain levels of confidence for the values being calculated. Most importantly, the tests have to be made based on real forensic casework, as the methods are heavily dependent on models estimated from real data. Testability must therefore not only hold on a theoretical level but must also include the very same data it will be applied on in forensic science.

Chapter 3

Methods

In this chapter the methods used for the experiment and the evaluation of results are described in detail. The following section deals with the parameters used for the speaker discrimination process. The subsequent sections provide insights into the calculation of the likelihood ratios, given the kind of multivariate data provided by the parameters, as well as methods to combine several scores into a single fused score. At the end of the chapter an outline is given of the methods used to evaluate the performance of the system as a whole, in terms of discriminatory power and calibration properties of the likelihood scores delivered by the system.

3.1 Formant feature extraction

The formant measurements which the method relies on are extracted from the signal by a formant-tracking algorithm. It is based on linear predictive coding (LPC) which is a method used extensively in digital speech processing. It makes use of a linear prediction (all-pole) model which rests upon the idea of estimating values in a discrete time series from the preceding output values (Markel & Gray 1976). The purpose of this procedure is to form a model of a vocal tract configuration during a given time frame as a linear time-invariant system.

$$\tilde{x}(n) = \sum_{i=1}^{q} \alpha_i x(n-i) \qquad\qquad \varepsilon(n) = x(n) - \tilde{x}(n) \tag{3.1}$$

The above equation shows the prediction of the present sample from the preceding output samples denoted by $x(n-i)$. ϵ represents the error made by the prediction with respect to the actual value of $x(n)$.

In the formant tracking procedure the speech input is presented as a discretised digital waveform and divided into frames. These frames containing the sampled values are preprocessed by a pre-emphasis filter and windowed. The samples are in turn used to arrive at a model of a specific vocal tract configuration, which is defined by the linear prediction coefficients α_i. For the estimation of the coefficients, the error ϵ is taken as a signal and is minimised. Methods for deriving a solution for the equations are the *covariance* and the *autocorrelation* methods (Markel & Gray 1976:166). The latter formulation produces a system of linear equation which can be solved by a very efficient recursive algorithm wherein each calculated coefficient is used for obtaining the following.

Following this procedure, initial formant estimates are obtained (*raw data*) which are in turn used by a tracking algorithm that fills the formant slots with the best candidate raw values to obtain the formant tracks.

This procedure represents the first basic feature extraction step applied in the method employed in the present work.

3.2 Parametric models of formant trajectories

This section deals with the processing of the formant tracks obtained by the aforementioned procedures to obtain final features for speaker discrimination that capture the speech dynamics in time of the uttered segments.

The method used for this study replicates the procedures outlined in Morrison & Kinoshita (2008) and Morrison (2009b) where it was applied to Australian English diphthongs.

Following the procedures used in these studies, parametric curves were fitted to each trajectory of the formant values extracted from the data. They used second and third order polynomials as well as discrete cosine transforms (DCT) to derive a parametric representation of the dynamic aspects of vocal tract movement during the production of the diphthong.

The coefficients of the parametric curves were then used as parameters in the likelihood ratio calculation using the multivariate kernel density formula developed by Aitken & Lucy (2004) which is described in the subsequent section.

The following sections give an in-depth description of the two types of representations.

3.2.1 Polynomial curves

A polynomial function is denoted by the sum of powers of its argument multiplied by coefficient values. Equation 3.2 shows the generic form of a polynomial function.

$$y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_k x^k \tag{3.2}$$

A polynomial of kth degree has $k+1$ degrees of freedom which are expressed as the $k+1$ coefficient values $\alpha_{0,\ldots,k}$ which determine the shape of the polynomial.

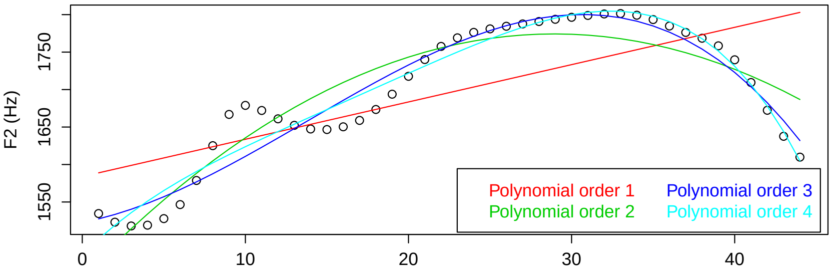


Figure 3.1: Polynomial fitting applied to formant data

Figure 3.1 shows the formant measurements of the second formant taken from one of the speakers in the corpus. The superimposed coloured lines represent polynomial functions fitted to the observed values, each using a different order.

The red line shows a polynomial of the form of a linear function $\alpha + \beta_1 x$, which is a rather poor approximation and captures little speaker-specific properties. The green and blue functions take the form of a quadratic and cubic function, respectively, where another term $\beta_i x^i, i \geq 2$ is added. These functions give a quite good account of the dynamic behaviour expressed in the formant trajectory without fitting too close to the particular observation.

The cyan curve shows the polynomial of fourth order. As has been shown in previous studies (McDougall & Nolan 2007), the use of this higher-order representation yields no gain in performance or even a decline, as it models the individual representations with too much detail, causing overfitting, and thus fails to provide a good generalisation of the speaker's time-dynamic speech properties.

Automatic fitting of polynomials

One method for fitting polynomials to a series of data points is commonly known as the *method of least squares*. Central to this procedure is the sum of squared residuals, which is derived by equation 3.3.

$$R^2 \equiv \sum_{i=1}^{n} [y_i - (\alpha_0 + \alpha_1 x_i + \ldots + \alpha_k x_i^k)]^2 \tag{3.3}$$

The best approximating polynomial is assumed to have the minimal sum of the deviations squared from the data points observed. For this the partial derivatives of the polynomial coefficients $\partial\alpha_0, \ldots, \partial\alpha_k$ must yield zero.

The present work uses the built-in functions for linear regression models within the R statistics software package (R Development Core Team 2009) to derive the polynomial coefficients from the individual formant trajectory measurements.

3.2.2 Discrete cosine transform (DCT)

The discrete cosine transform (DCT) uses the sum of cosine functions with different frequencies and amplitudes to express a finite set of data points descending from some arbitrary function or a digitalised signal. It is used in modern image compression algorithms. The principle advantage of the DCT is to remove redundancy in the data, leading to decorrelated transform coefficients.

Equation 3.4 shows the formula to construct a DCT curve from the coefficients (*inverse DCT*).

$$y(x) = \alpha_0 \frac{1}{\sqrt{N}} + \sum_{k=1}^{K} \alpha_k \frac{2}{\sqrt{N}} cos\left(\frac{(2x+1)\pi k}{2N}\right) \tag{3.4}$$

The α_k represent the coefficient values, N is the number of data points, and K denotes the order of the curve. Each coefficient represents one cosine base function. The first produces a straight line at the height of the mean of the data points, the second a cosine of a half cycle which gives the direction and the magnitude of the formant trajectory's tilt, and the third cosine captures the curvature of the trajectory. A prior study by Watson & Harrington (1999) investigated the use of DCT coefficients to model the dynamic behaviour of formants for classifying vowel and diphthong phonemes, which concluded that monophthongal vowels could be discriminated by static targets alone while diphthongs required more dynamic information.

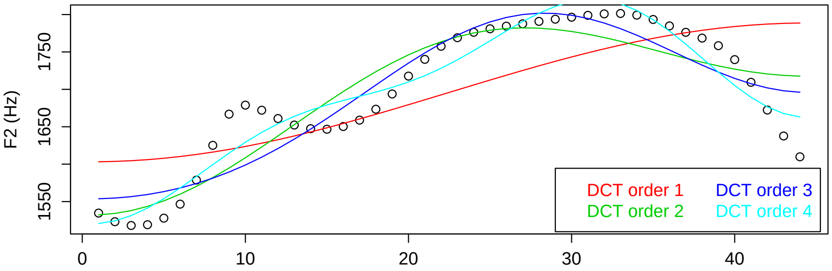


Figure 3.2: DCT applied to formant data

Figure 3.2 shows the same formant trajectory as is shown in the polynomial fitting example (figure 3.1), with curves constructed by using inverse DCTs of different orders. The red curve is obtained by applying inverse DCT with only two coefficients (α_0 and α_1), yielding a cosine of a half cycle. The other curves of higher degrees are created the same way, but with additional coefficients. As can be seen by comparing the polynomial and DCT curves they show a rather similar behaviour in respect to how well they align to the trajectory with increasing order of the curve.

Morrison & Kinoshita (2008), Morrison (2009b) used this kind of parametric representation in their studies and concluded that '[t]here were trends indicating that DCTs generally outperformed polynomials [. . .]'. However, this observation depends very much on the data which the methods are applied to and it is expected to find variability in performance between different diphthong segments of other language varieties.

Calculation of discrete cosine transform coefficients

The discrete cosine transform was calculated using the package `dtt`[1] (Komsta 2007) for the R statistics software package (R Development Core Team 2009), which provides functions for several discrete trigonometric transformations.

[1] `http://cran.r-project.org/web/packages/dtt/dtt.pdf` (retrieved 2009-05-12)

The vector of measurements is transformed into its DCT components. Of these, the coefficients three to four are used as parameters in tests denoted as DCT order 2 or 3, respectively. This is done in analogy to the polynomial functions of different degrees which use the same number of coefficients as features.

3.3 Likelihood ratio calculation

As noted in section 2.3.2 the outcome of the approach outlined in this thesis should ideally be expressed as a likelihood ratio which gives an assessment of the strength of evidence for use in the Bayesian approach.

An analytic formula for obtaining likelihood ratios from continuous multivariate data has been developed and described in Aitken & Lucy (2004). It assesses the difference between the samples taken from the suspect and the offender sample with respect to a background distribution estimated from a given population.

$$\frac{\begin{array}{c}(2\pi)^{-p}|D_1|^{-\frac{1}{2}}|D_2|^{-\frac{1}{2}}|C|^{-\frac{1}{2}}(mh^p)^{-1}|D_1^{-1}+D_2^{-1}+(h^2C)^{-1}|^{-\frac{1}{2}} \\ \times\exp\{-\frac{1}{2}(\mathbf{\bar{y}_1}-\mathbf{\bar{y}_2})^T(D_1+D_2)^{-1}(\mathbf{\bar{y}_1}-\mathbf{\bar{y}_2})\} \\ \times\sum_{i=1}^{m}\exp[-\frac{1}{2}(\mathbf{y}^*-\mathbf{\bar{x}_1})^T\{(D_1^{-1}+D_2^{-1})^{-1}+(h^2C)\}^{-1}(\mathbf{y}^*-\mathbf{\bar{x}_1})]\end{array}}{\begin{array}{c}(2\pi)^{-p}|C|^{-1}(mh^p)^{-2}\prod_{l=1}^{2}[|D_l|^{-\frac{1}{2}}|D_l^{-1}+(h^2C)^{-1}|^{-\frac{1}{2}} \\ \times\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\mathbf{\bar{y}_l}-\mathbf{\bar{x}_i})^T(D_l+h^2C)^{-1}(\mathbf{\bar{y}_l}-\mathbf{\bar{x}_i})\}]\end{array}} \tag{3.5}$$

$$\mathbf{y}^*=(D_1^{-1}+D_2^{-1})^{-1}(D_1^{-1}\mathbf{\bar{y}_1}+D_2^{-1}\mathbf{\bar{y}_2})^{-1} \tag{3.6}$$

$$h=\left(\frac{4}{2p+1}\right)^{1/(p+4)}m^{-1/(p+4)} \tag{3.7}$$

Equation 3.5 shows the likelihood ratio multivariate kernel density formula as it was presented in Aitken & Lucy (2004). A detailed characterisation is given hereinafter.

The population of p characteristics of items, that is the set of parameters taken from realisations of a diphthong, is denoted as Ω. The background data is taken as a random sample of m members from Ω, with n measurements of the characteristics each, and is labelled as $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T, i = 1, \ldots, m, j = 1, \ldots, n$. The total number of measurements is denoted by $N = nm$.

The suspect and offender measurements are denoted by $\{\mathbf{y}_l\} = (\mathbf{y}_{lj}, j = 1, \ldots, n_l, l = 1, 2)$, where $\mathbf{y}_{lj} = (y_{lj1}, \ldots, y_{ljp})^T$. Their distributions conditional on the source are assumed to be normal, with the theoretical mean $\theta_l, l = 1, 2$ and variance-covariance matrix $D_l, l = 1, 2$.

The Gaussian distribution's parametric nature enables its use on quantitatively rather limited data which is the case commonly faced in forensic cases, provided that the parameters in fact follow a normal distribution. As with parameters used in speaker recognition, trace evidence displays within-source and between-source variation. The likelihood ratio formula takes this into account by deriving statistical models from the data given to calculate a score.

The within-speaker variance is also modelled by a Gaussian distribution with the theoretical mean θ_i, estimated from the measurements $\{x_{ij}\}$ for speaker i, and the within-speaker variance-covariance matrix U, which is estimated from the background data as follows.

$$\hat{U} = \frac{S_w}{N - m} \qquad\qquad S_w = \sum_{i=1}^{m} \sum_{j=1}^{n} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \tag{3.8}$$

The between-speaker variance models the distribution of the within-speaker theoretical means θ_i. This distribution is not necessarily normal. Therefore, the formula described in Aitken & Lucy (2004) uses a kernel distribution for modelling between-speaker variability. In this technique a probability density function is estimated by taking the sum of Gaussian functions for each observation, with its parameters mean and variance set to the observed value and a smoothing factor, respectively, normalized by the number of observations.

Figure 3.3 shows a histogram of non-normally distributed data and a kernel density estimation which provides a better fit than a Gaussian distribution. The advantage of a kernel density estimate lies in its non-parametric nature which allows a better approximation of the data. Given a representative dataset of the background population, the actual distribution of features can be more accurately

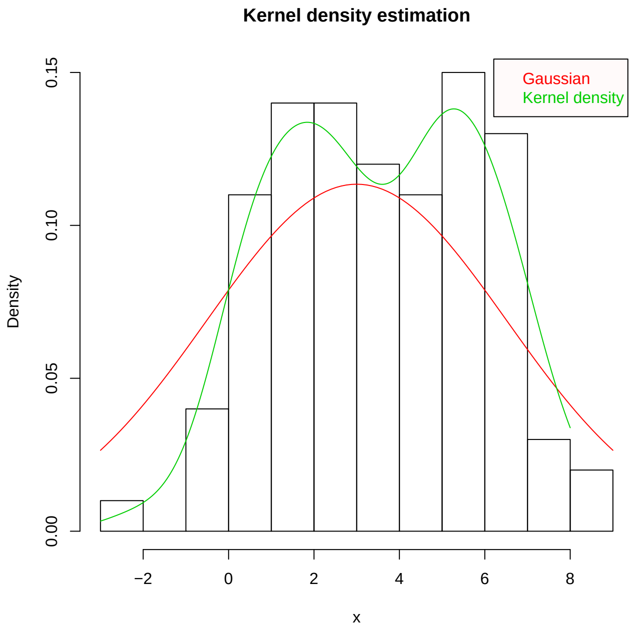


Figure 3.3: Example of kernel density estimation

modelled. Furthermore, as a rather technical convenience, it is always guaranteed to be a proper probability density function, i.e. is non-negative and integrates to one.

For modelling the between-speaker variability a multivariate normal density function is used as the kernel density function. The parameters to this function are the empirical within-speaker means $\bar{\mathbf{x}}_i$ and the covariance matrix h^2C which is detailed below.

$$\hat{C} = \frac{S^*}{m-1} - \frac{S_w}{n(N-m)} \qquad S^* = \sum_{i=1}^{m} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \tag{3.9}$$

The estimation of the smoothing parameter h was declared in equation 3.7. The overall probability density function for the between-speaker variance is given in the following equation.

$$f(\theta|\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_m, C, h) = \frac{1}{m} \sum_{i=1}^{m} K(\theta|\bar{\mathbf{x}}_i, C, h) \tag{3.10}$$

The parameters for the Gaussian distribution of the suspect and the offender models are estimated from the data. The theoretical means θ_l are estimated from the measurements $\mathbf{y}_l$, but the variance-covariance matrices $D_l, l = 1, 2$ are adapted from the within-group covariance matrix U by division with the number of measurements used for the suspect or offender.

The formula's original formulation was intended for the quantitative numerical evaluation of trace evidence in form of glass fragments found at a crime scene and on a suspect (Aitken & Lucy 2004), but it has subsequently been shown to be applicable in the domain of speech evidence, using formant measurements as multivariate data (Rose 2005). This formula has successfully been applied to forensic phonetic data (see for example Rose et al. (2006), Morrison (2008, 2009b), Morrison & Kinoshita (2008)).

The multivariate data used in the method outlined in this chapter is comprised of the coefficients derived by approximating the polynomial functions to the formant trajectories as well as using the first three or four DCT coefficients, as outlined in the preceding sections.

3.4 Evaluation of performance

The question of how to evaluate and compare the performance of recognition and classification systems in general is by far not trivial. For the task of speaker recognition, several metrics and representations have been developed. These include the equal error rate (EER) as a single-number assessment of performance, the detection error trade-off (DET) plot as a comprehensive summary representation of discrimination performance, the Tippett plot and the Applied Probability of Error (APE) plot along with the log likelihood ratio cost metric (C_{llr}) which quantify the loss in performance due to discrimination errors and the calibration of the system.

To evaluate recognition systems, a series of *trials* are performed in which the system under evaluation must give either a categorial decision, i.e. it must determine whether the speech recorded in the training (suspect) sample and the test (offender) sample originate from the same speaker, or a score indicating a "strength of belief" of speaker similarity or dissimilarity. These trials can be categorised into the following two groups.

- *target trials*, where the target speaker is indeed the speaker on the test sample.

- *impostor trials*, where different speakers are involved.

The system should ideally give a positive response in the former case and a negative in the latter, which can take the form of a likelihood ratio score as a measure of strength of the decision or a categorial accept/reject answer which is often derived by setting a threshold on the score scale.

The performance is then assessed by measuring the decision errors made by the system. The two errors generated are *missed decisions*, meaning that a target trial has been rejected as not originating from the same speaker, and *false alarms*, where the system returns a positive result for a impostor trial. The probabilities of error corresponding to these errors are denoted by $P_{miss|target}$ and $P_{FA|Impostor}$, respectively (Przybocki et al. 2007).

Figure 3.4 shows a plot of the distributions of log likelihood ratios typically assigned to speaker detection trials. The target trial distribution overlaps with the impostor trial distribution. The definition of a decision threshold yields the two error rates described above.

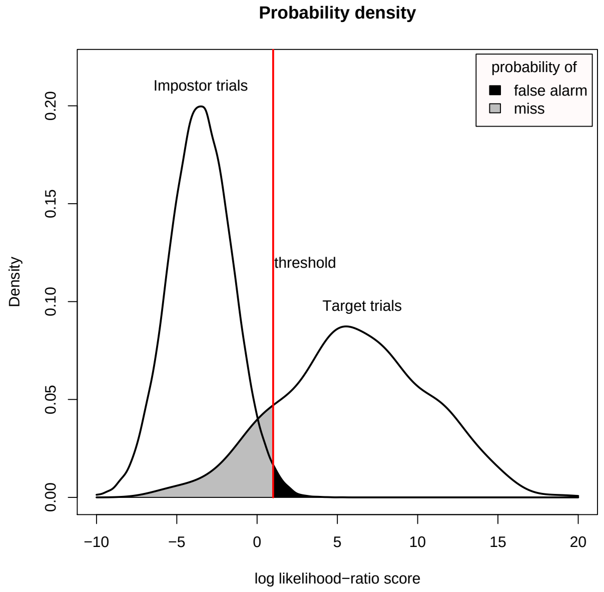


Figure 3.4: The log likelihood ratio distributions for target and impostor trials (after van Leeuwen & Brümmer (2007:334))

The evaluation metrics listed at the beginning of this section have been incorporated into the evaluation plan of the last NIST[2] speaker recognition evaluation (SRE) in 2008 (NIST 2008) and have been widely adopted in the field of automatic speaker recognition. For the reason of comparability to other approaches to (forensic) speaker recognition systems these measures are adopted in this thesis. The subsequent sections give a detailed account of their evaluation focus and capability.

3.4.1 Discrimination

The performance of a system is preeminently assessed in terms of its discriminatory potential, i.e. how well it can discriminate between speakers. This property is evaluated by comparing error rates for the two detection errors described in the previous section. However, as there are two types of errors involved, a comparison between systems cannot be achieved straightforwardly, because the two errors are in a trade-off relation to each other, as the number of misses naturally increases and the number of false alarms decreases as the decision threshold moves upwards the LR scale, and vice-versa.

3.4.2 Calibration

Calibration refers to the relation between the information gained by the output of a system and its interpretation by the *finder-of-fact*, i.e. the judge or jury (Ramos 2007:99). The decision of the court is made on a statement of the strength of evidence, either in favour or against the defendant. Therefore, it should be presented in a consistent and precise way that enables a straightforward interpretation.

The calibration properties of a system describe how well-aligned the output scores are with respect to a scale that is used for interpretation by the finder-of-fact. In case of likelihood ratios, the scale applied is the one described in section 2.3.2 where values greater than one support the prosecution hypothesis while values less than one give support for the hypothesis of the defence (see 2.2).

The score produced by a system has no absolute meaning in itself, even if they approximate the likelihood ratio scale. Values can get indefinitely large or small. To produce values that adhere to a consistent scale the scores can be scaled and

[2] National Institute of Science and Technology

shifted. This will not affect discrimination performance, as the decision threshold is shifted with the scores (van Leeuwen & Brümmer 2007:339).

3.4.3 Equal Error Rate (EER)

The EER gives a characterisation of a system as a single value which can be directly used for comparison. Since, as outlined above, the error rates are in a trade-off relation, there is a point where they are equal. This value is called the Equal Error Rate or EER.

This figure is used to compare the discrimination performance and is independent of the scale of the scores which are produced by the system, meaning that the value can be the same for systems that yield a likelihood ratio as for ones that deliver some form of distance scores as it is depends only on the number of matches and mismatches of the two types of trials.

Being a single value, this measure does not provide sufficient information with regard to the scores output by the system as well as their distribution. Therefore, other tools must be additionally provided that give a more detailed picture at different operation points.

3.4.4 The Detection Error Trade-off (DET) plot

The Detection Error Trade-off plot (Martin et al. 1997) is a graphical representation of the inherent trade-off between the two error types. It gives a characterisation of a system over the whole range of possible decision threshold values. Figure 3.5 gives an example based on scores from a speaker recognition system.

The axes are warped according to the quantile function of the Gaussian distribution. In figure 3.4 the distributions of the likelihood ratios were presented along with the threshold that yielded the trade-off between the two types of errors. In the DET plot, instead of plotting the miss and false alarm probabilities, the standard deviations corresponding to these probabilities are plotted (Martin et al. 1997:2). The scales of the standard deviation are plotted on the top and left of the figure.

As a consequence the DET plot of a system approximates a straight line if the two error rates are normally distributed. The slope of the line depends on the ratio of the standard deviations of the impostor and target trial distributions

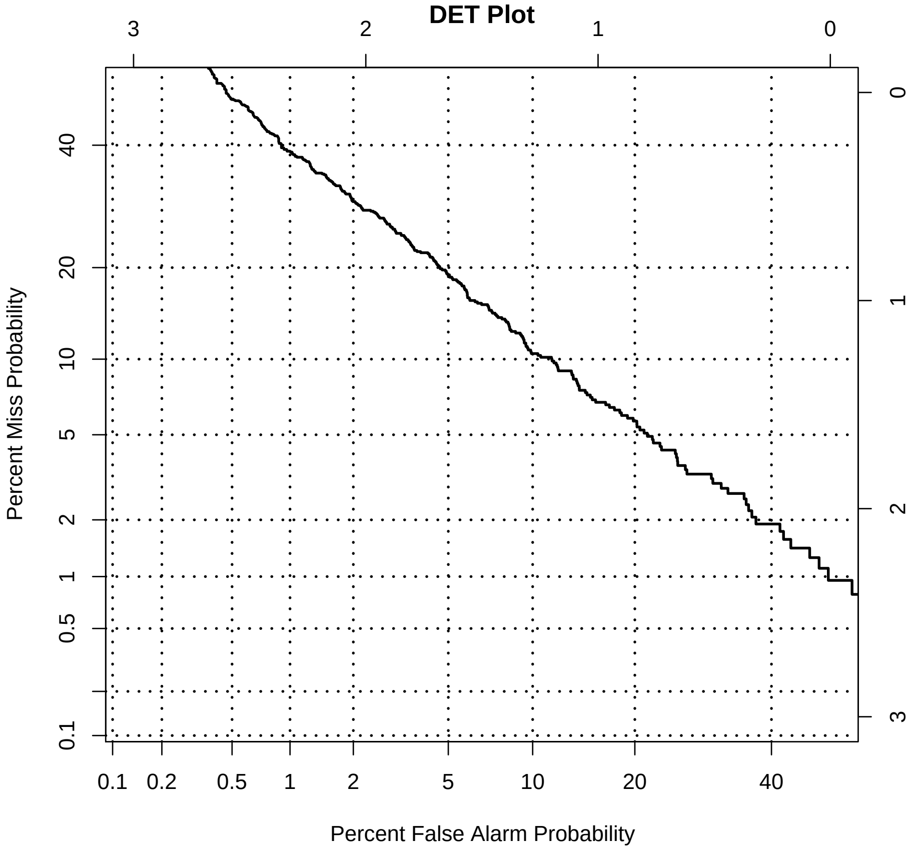


Figure 3.5: Example of a Detection Error Trade-off (DET) plot

(van Leeuwen & Brümmer 2007:334). Random performance is indicated as a straight diagonal line at $y = -x$. The detection error trade-off plot shows only the lower left quartile because the performance of the systems under comparison is usually better than random.

A main property of the DET plot is that the performance of several methods for speaker identification can be easily compared in one figure. Better performance is indicated by curves further to the lower left and even small improvements are easily perceived. It also has to be noted that, although thresholds are used to calculate the error rates, the performance expressed by the DET plot can be assessed without the need that the systems compared actually involve the setting of a threshold or arriving at a categorial decision. This is important for the use of the likelihood ratio as a probability statement of the strength of evidence, as has been mentioned in chapter 2.3.2.

As with the EER the DET plot is a measure of discrimination performance based on error rates and is not bound to scores that adhere to the likelihood ratio scale.

3.4.5 The Tippett plot

Tippett plots are a graphical representation of $P_{miss|target}$ and $P_{FA|Impostor}$ as a function of the (log) likelihood ratio. In the plot two curves are displayed that indicate the probability for the respective hypotheses, $H0$ representing the hypothesis of the prosecution and $H1$ expressing the competing hypothesis of the defence, given a log likelihood ratio score. The name refers 'to the concepts of "within-source comparison" and "between-source comparison" defined by Tippett (1968)' (Alexander et al. 2004:97). An example is given in figure 3.6.

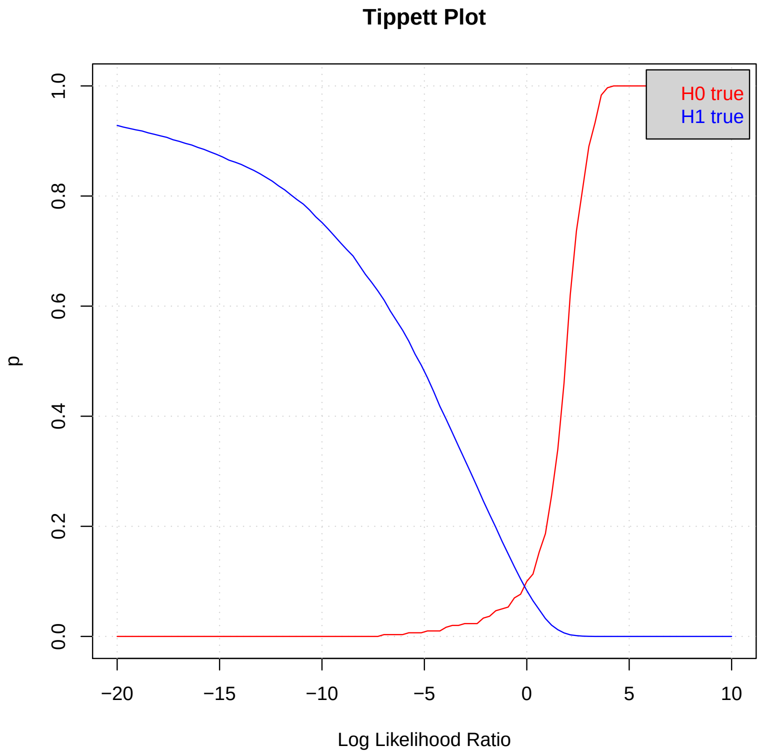


Figure 3.6: Example of a Tippett plot

The red curve of the prosecution hypothesis represents the distribution of the scores calculated for target-speaker trials and the blue curve of the hypothesis of the defence shows the distribution of non-target-trial scores. They represent the proportion of likelihood ratios greater than a given likelihood ratio for each hypothesis.

The Tippett plot provides a solid graphical evaluation tool for the calibration properties of a system, as the curves representing the respective hypotheses will saturate much faster for likelihood ratios deviating from zero if the scores are well-calibrated. It is a measure of quality for the scores produced by the system during a test, as the meaning inherent to the scores and their distribution is important in order to be able to state how certain it is that the same speaker was involved if the system reports a specific score. For example, in the Tippett plot in figure 3.6 one can be to almost 90% sure that if the log likelihood ratio calculated by the system is -15 the offending speech sample was produced by a different speaker. It is however important to keep in mind that this is only applicable to scores produced in a test and the example given is not extendible to future scores, unless the conditions of the training and the real case data are comparable.

3.4.6 The log likelihood ratio cost function C_{llr}

The C_{llr} function has been introduced by Brümmer & du Preez (2006) to provide a metric that simultaneously measures discrimination and calibration performance. Like the Tippett plot it is a quality measure for speaker detector scores, but particularly for values adhering to the log likelihood ratio scale.

Prior to its inception, the Decision Cost Function C_{det} was used for the same purpose which required specifying the prior probability of targets (see section 2.3.1) and the costs of $P_{miss|target}$ and $P_{FA|Impostor}$ errors as application-dependent parameters (van Leeuwen & Brümmer 2007:337).

The scores produced by a speaker classifier can spread over any range and can be scaled and shifted accordingly. As van Leeuwen & Brümmer (2007:339) notes '[t]here is no meaning in the scores, other than an ordering'.

The C_{llr} metric represents a function that attaches costs to log likelihood ratios based on their position on the likelihood ratio scale, which in principle ranges from negative to positive infinity with zero as a threshold. The basic assumption is that target-trials should yield high LLR values and non-target-trials should produce low (i.e. negative) LLR values. Deviations from this concept are 'punished' with a higher cost.

The log likelihood ratio cost function 'sample[s] C_{det} over an infinite "spectrum" of operating points and then to simply integrate over them' (van Leeuwen & Brümmer 2007:341). This makes the error probabilities a function of the threshold and thus represent the information provided by the Decision Cost Function over the whole range of thresholds.

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}} log_2(1+\frac{1}{LR_{ss_i}}) + \frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}} log_2(1+LR_{ds_j})\right) \qquad (3.11)$$

Equation 3.11 shows the analytical form of the C_{llr} metric as it is presented in Morrison & Kinoshita (2008). N_{ss} and N_{ds} are the number of target and impostor trials. LR_{ss} and LR_{ds} are the likelihood ratios of the trials, respectively.

The value for C_{llr} for a system that returns infinite LR values for target-trials and zero for non-target-trials would be zero, meaning a perfect system. However, a system that produces same-speaker likelihood ratios close to one or even lower is qualified with a high value.

To assess the calibration properties of a system the C_{llr}^{min} value is calculated which is the minimum loss possible if the system were optimally calibrated and is therefore a measure of discrimination. C_{llr}^{cal} is the the calibration loss which is the difference between C_{llr} and C_{llr}^{min}.

The C_{llr}^{min} value is assessed by deriving values for a monotonic rising warping function w which scales and shifts the likelihood ratio values output by the system. This warping function is derived by applying the Pool Adjacent Violators (PAV) algorithm (Brümmer 2004). The C_{llr} value calculated for likelihood ratio values after applying this algorithm constitutes C_{llr}^{min}.

3.4.7 The Applied Probability of Error (APE) plot

The Applied Probability of Error (APE) plot is a graphical representation of the error probability over the range of possible operating points, i.e. thresholds. In the Bayesian framework these are represented by the prior odds (see section 2.3.1).

This error probability is given by the following formula.

$$P_e(\theta) = \tilde{P}_{tar}(\theta)P_{miss}(\theta) + (1 - \tilde{P}_{tar}(\theta))P_{FA}(\theta) \tag{3.12}$$

P_e is a function of the prior log odds, i.e. the logarithm of the prior probability in odds form. The APE plot graphs this function against an interval of the logit prior θ close to zero.

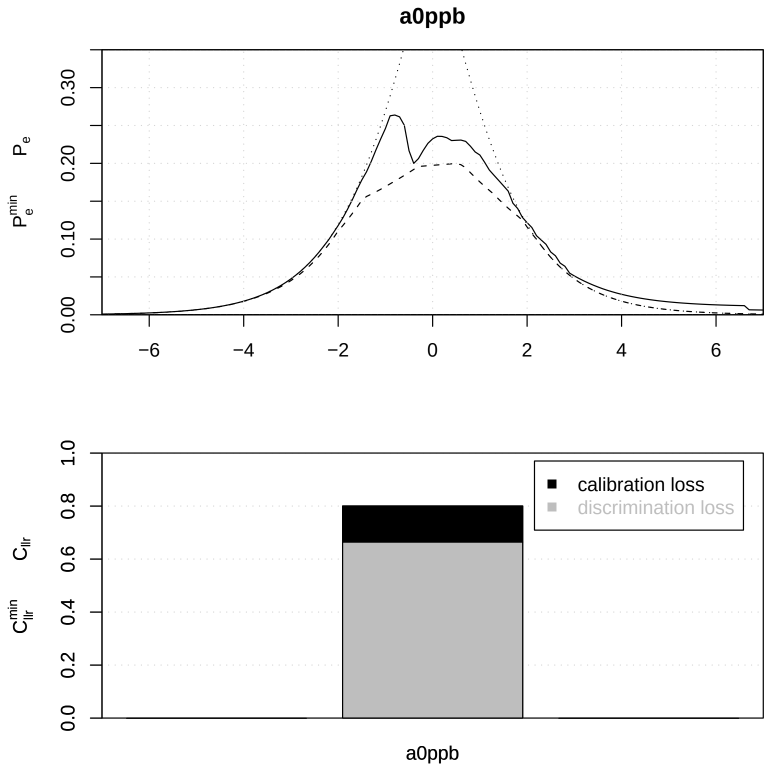


Figure 3.7: Example of an Applied Probability of Error (APE) plot

Figure 3.7 shows an example of an APE plot. The horizontal axis represents the logit prior and the vertical axis shows the error probability.

Three curves are plotted which represent three kinds of information.

- The solid curve represents the error probability of the system under evaluation. The area beneath is proportional to the log likelihood ratio cost function C_{llr}.
- The dotted curve represents the reference detector which always returns a likelihood ratio of 1 and, thus, the decision is based solely on the prior.

This is useful as it shows that, if the solid curve crosses the dotted line the system gives more decremental information than a system that does not incorporate the speech data into its decision. For a logit prior of zero, i.e. a prior probability of one, the probability of error for this reference detector is 50%, hence chance level.

- The dashed curve shows the probability of the system with the warping function applied to its output scores. The area beneath is proportional to the discrimination loss C_{llr}^{min}.

The lower part is essentially a bar plot of the C_{llr} values of the systems under comparison. The grey area represents the discrimination loss, i.e. C_{llr}^{min}, and the black portion is the loss of information ascribed to less-than-optimal calibration.

The APE plot is used in the evaluation to show both the discrimination properties as well as the calibration properties of the method used in this thesis.

3.5 Post-hoc calibration of likelihood ratio scores

In automatic speaker recognition systems there usually exists a calibration stage that transforms the output difference scores into well-behaved likelihood ratios, meaning that the resulting values are in fact aligned to the likelihood ratio scale in that higher values are returned indicating stronger evidence in favour of the prosecution hypothesis and lower values are returned indicating stronger evidence against it (see section 3.4.2).

As Morrison & Kinoshita (2008:1502) state '[t]he aim of calibration [...] is to present the information in such a way as to best aid the finder of fact in making appropriate decisions'. This is done to ensure that the score returned by the system is more consistent and more easily interpretable in court.

As has been shown by Morrison & Kinoshita (2008) and Morrison (2009b) this procedure can also be applied to likelihood ratios obtained by the multivariate kernel density formula described above (see section 3.3).

Calibration can be attained either by defining a fixed function that warps output scores into likelihood ratios or by using discriminative methods that optimise a given objective function.

One representing the latter is the *S-cal* method which performs linear mapping, i.e. shifting and scaling, as well as sigmoid saturation step. This is performed by the following equation.

$$llr(s) = \log \frac{(logit^{-1}\alpha)(e^{a \cdot s + b} + 1) + 1}{(logit^{-1}\beta)(e^{a \cdot s + b} - 1) + 1} \tag{3.13}$$

Following the formula, the score (represented by s) is scaled and shifted by the parameters a and b. The effect of the subsequent sigmoid saturation depends on α and β, which leads to monotonically increasing mapping if $\alpha > \beta$ or decreasing mapping if $\alpha < \beta$. If α is much greater than zero and β is much lower than zero then the minimum and maximum of the resulting monotonically increasing sigmoid are defined in terms of α and β. If, however, both α and β are equal then the resulting score is zero.

These conditions are best described graphically. Figure 3.8 shows the effect of calibration by S-cal using two different sets of parameters. The curve has a sigmoid form, hence the initial letter *S*. As can be seen the sigmoid form yields calibrated scores that are saturated at both extremes.

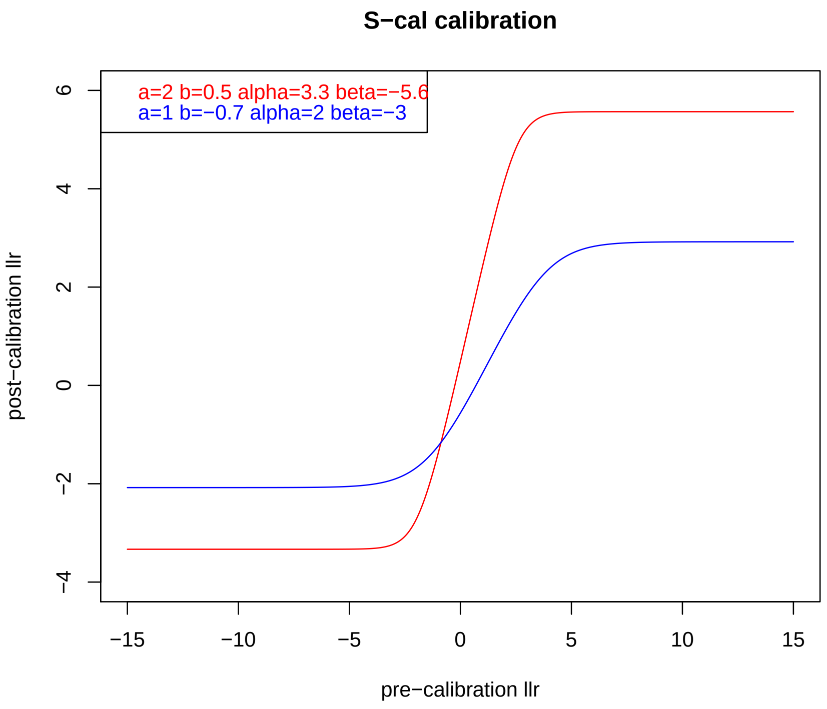


Figure 3.8: Effect of S-cal calibration on log likelihood ratios

The parameters needed for calibration are obtained by training on development scores of supervised calibration targets, which must be carefully selected using the following criteria.

- The selected trials should be as similar to real targets that the calibration will be applied in the future.
- The number of trials must be sufficiently large
- The trials should not be used for other system-wide training, such as for fusion weights (see section 3.6).

The training methods are included in the *FoCaL* toolkit (Brümmer & du Preez 2006) which perform numerical optimisation on the scores of target and non-target calibration trials to minimise the C_{llr} metric (see section 3.4.6).

3.6 Fusion approaches for combining likelihood ratio scores

The term *fusion* refers to applying a function on likelihood ratios or, more general, real-valued scores of any kind supplied by several different systems to arrive at a single likelihood ratio for a system which incorporates the results of all individual systems.

Likelihood ratio scores can also be combined by multiplying their output scores. Due to the fact that the scores produced are possibly highly correlated, this often results in very extreme combined likelihood ratio scores that massively overstate the strength of the evidence, leading rather to confusion than easily interpretable scores.

The form of the fusion function can take many different forms. One is the use of linear combination of the individual scores which is employed in this thesis.

$$s_f = s(x, \mathbf{w}) = w_0 + \sum_{i=0}^{N} w_i s_i(x) \tag{3.14}$$

Equation 3.14 describes the principle of linear fusion in which the fused score is the sum of scores, e.g. log likelihood ratios, s_i weighted by fusion weights w_i.

The individual scores can be scaled and shifted to yield a resulting score that shows better discriminatory potential and is properly calibrated, i.e. allows for easy interpretation of the output. The fusion weights must be trained on supervised likelihood ratio scores given a function that must be optimised against during the training stage. Afterwards they can be used to fuse new likelihood ratio results by the same systems.

These training scores must be available from all systems whose outputs should be fused. The underlying training trials must be selected carefully, as they define how future likelihood ratios should be scaled and shifted, and thus modified, without actual reference to the data at hand.

The procedure described is called *linear logistic regression fusion* which is implemented within the *FoCaL* toolkit (Brümmer & du Preez 2006). The name originates from the logistic regression objective which must be optimised against. This objective is stated in equation 3.15.

$$\begin{aligned} C_{wlr} = & \frac{P_{tar}}{\|\chi_{tar}\|} \sum_{x \in \chi_{tar}} \log(1 + e^{-s(x,\mathbf{w}) - logit P_{tar}}) \\ & + \frac{1 - P_{tar}}{\|\chi_{non}\|} \sum_{x \in \chi_{non}} \log(1 + e^{s(x,\mathbf{w}) + logit P_{tar}}) \end{aligned} \tag{3.15}$$

This objective represents a cost function which must be minimised. P_{tar} represents a given prior which, if set to 0.5, results in the objective to resemble the C_{llr} metric, which is used for evaluating calibration performance (see section 3.4.6).

The function is convex, meaning that, pictorially, it lies below a straight line connecting any two points of the function, which results to it having only one global minimum. Within the *FoCaL* toolkit, a conjugate gradient algorithm is used for finding this minimum.

Chapter 4

Experiment and results

This chapter presents the experimental setup and results based on the method outlined in the previous chapter. The first section presents the data corpus on which the inquiry is based. The subsequent sections deal with the evaluation of the results in terms of the discriminatory potential and the calibration properties of the method.

4.1 Corpus of Austrian German - OeD

The data used for this study consists of recordings of 30 male speakers of Viennese Austrian German, aged 20 to 70, which has been collected over several years at the *Acoustics Research Institute (ARI)*[1]. The speakers were recorded while performing several tasks:

- Repeat sentences (standard and dialect variety)
- Reading standard and dialectal texts
- Spontaneous speech

The data used for this experiment was taken from the word *kreidebleich* /ˈkraɛdɛˌblaɛɕ/ ('chalk-white') in the following repeated (standard variety) sentence.

(1) Nach einer Feier liegen alle kreidebleich am Boden
After a party everybody is lying on the floor, white as chalk

[1] `http://www.kfs.oeaw.ac.at/`

The word contains the diphthong /aɛ/ in a primary and a secondary stressed position. Within the analysis they were treated separately, as they differ in stress and phonetic context. Each speaker was recorded while repeating this sentence ten times.

4.1.1 Motivations for using Viennese German /aɛ/

The first point to mention regarding the motivation of this study is that the time-dynamic properties of diphthongs encode a fair amount of speaker specific characteristics. Several studies have dealt with diphthongs and studied different kinds of representations for capturing this information (see section 2.2.6).

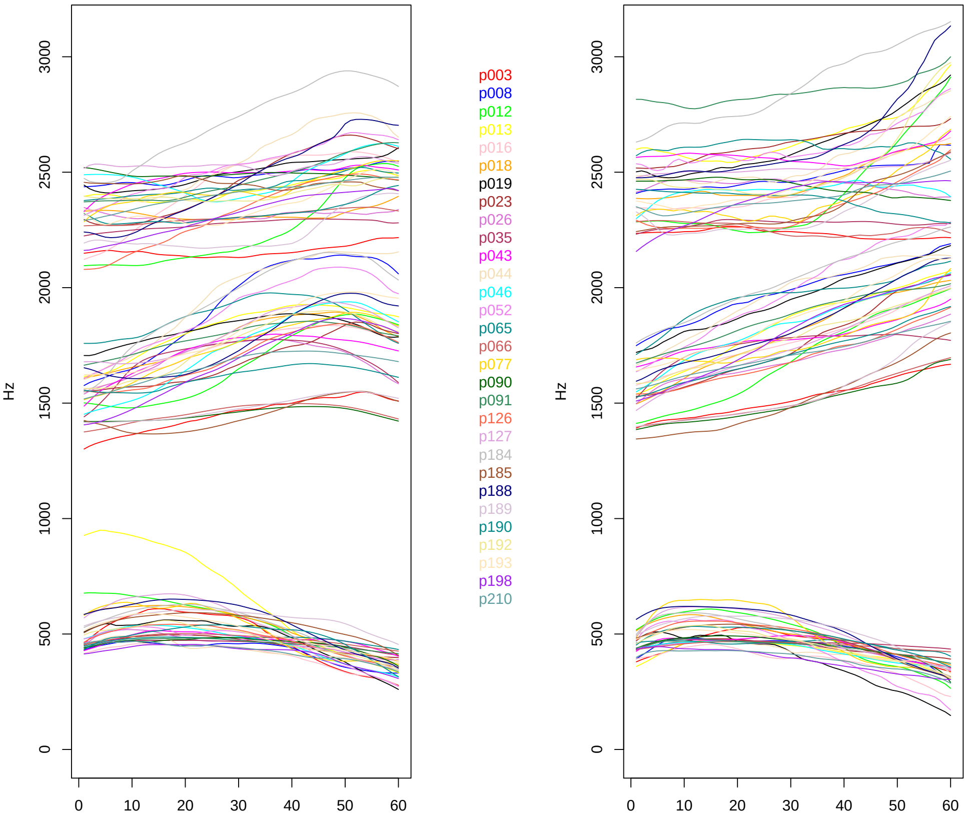
The language specificity of these characteristics, e.g. differences in relative timing and duration of onset, glide and offset sub-segments, requires investigation of the actual performance of methods proposed for discriminating speakers. As the method used in this study has still to be tested in terms of applicability on the diphthongs and vowels of other languages than the ones that were examined in previous papers, the present study provides insight into its use on Viennese German diphthongs.

Additionally, the effects of stress position and phonetic context on the extracted parameters that this method depends on require further study, as the dynamics of unstressed or secondary stressed diphthongs as well as their duration are clearly reduced. Thus, two different contexts and stress positions were chosen from the data available to acquire data about the performance of this method applied under these conditions.

4.1.2 Between- and within-speaker variation exemplified by Viennese diphthong dynamics

To provide insight into the range and extent of the variation between as well as within speakers this section will provide a characterisation of the dynamic differences exhibited by the speakers in terms of formant frequency trajectories.

Figure 4.1 provides a visualisation of time-equalised mean formant trajectories. As can be seen the trajectory of the first formant shows less between-speaker variability than the other two formants. Furthermore, the second and third formant display much diversity in the form of the trajectory contour.



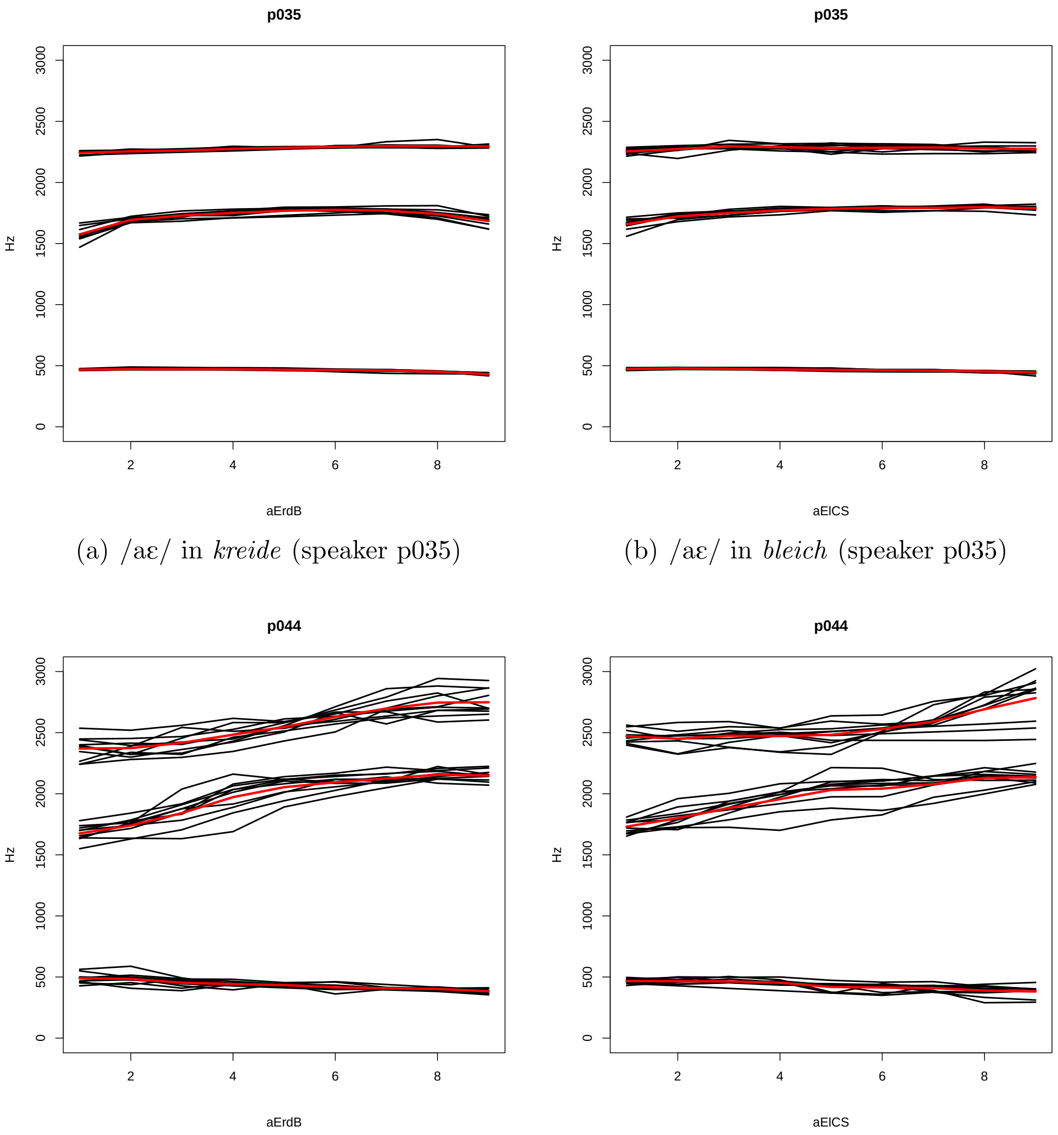
(a) /aɛ/ in *kreide*

(b) /aɛ/ in *bleich*

Figure 4.1: Time-equalised mean formant trajectories of the 30 speakers

Differences between the two segments can be readily recognised, again especially in the movement of the second and third formants. F_2 displays a rather gliding shift in the secondary stressed /aɛ/ in *bleich* while more genuine diphthongal properties can be located in the trajectory of the primary stressed /aɛ/ in *kreide*.

The extent and difference in magnitude of within-speaker variability can be seen in figure 4.2. Here the formant trajectories of two speakers are displayed, along with their mean frequency trajectory. Speaker *p044* exhibits rather large variability, whereas speaker *p035* shows only slight differences in the contours of the individual formant progressions. This is of particular importance with reference to the fact that speakers' formant trajectories are to be modelled by Gaussian distributions parametrised by the empirical sample mean and variance



(a) /aɛ/ in *kreide* (speaker p035)

(b) /aɛ/ in *bleich* (speaker p035)

(c) /aɛ/ in *kreide* (speaker p044)

(d) /aɛ/ in *bleich* (speaker p044)

Figure 4.2: Formant trajectories of individual utterances of two speakers

of the respective features.

4.2 Experimental setup

This section presents the experimental setup which includes the list of methods used for comparison as well as a detailed account of tasks that where defined for this purpose.

The following methods were applied to the data corpus to allow for comparison which are described below.

- Arithmetic mean of the formant measurements in the trajectory
- 10% measurement intervals (McDougall 2005, 2006)
- Dual-Target model, using relative (but fixed) time points within the trajectory (at 10% and 90% of the segment length)
- Fitting of polynomials of second and third degree (see section 3.2.1)
- Discrete cosine transform (DCT), using second and third order curves (see section 3.2.2)

To perform the speaker discrimination tests, the methods for approximating parametric functions to the formant trajectories and for the evaluation of the system performance that have been outlined in chapter 3 were implemented as computer programs using the R statistics package (R Development Core Team 2009)[2].

To enable a comparison between the different methods several different tasks were defined, which can be categorised by means of the number of features they incorporate.

Trials were run using data either from the first three formants or from only the second and the third. This distinction has been made because, as noted before, it is very often the case in forensic phonetics that recordings have been taken from telephone conversations, which are band-pass filtered at approximately 300-3400 Hz due to technical reasons. The first formant is often affected by this filtering, especially in vowels and diphthongs with very low F1, rendering it unusable as a feature. Thus, separate trials have been performed on both sets of data to assess the loss of speaker discriminating information by discarding F1, denoted as **f1-2-3** for all three formant measurements, and **f2-3** when using only the second and the third formant.

The following trials were performed along these dimensions. At the end of each condition the number of included parameters is given in parentheses.

[2] The evaluation procedures provided by the *FoCaL* toolkit (Brümmer & du Preez 2006) were ported from MATLAB to R by Timo Becker.

- **Formant means**

  The following tasks utilise the mean of the formant values in the trajectory as features

  f1-2-3_mean Incorporates the mean of all three formant trajectories (3 p.)

  f2-3_mean Incorporates the mean of the trajectories of the second and third formant (2 p.)

Secondly, to capture the differences in overall duration of the speakers, the raw segment length in seconds was used as an additional explicit duration feature. Trials incorporating this parameter are tagged with the identifier **dur**. The following tasks were performed and evaluated twice, with and without explicit duration information.

- **Interval and target measurements**

  In the following tasks measurements at relative targets or intervals are taken of each of the individual segments.

  f1-2-3_int10 Formant measurements taken at 10% intervals using all three formants (27 p.)

  f1-2-3_target Formant measurements taken at relative targets (at the 10% and 90% measurement) using all three formants (6 p.)

  f2-3_int10 Formant measurements taken at 10% intervals using formants F_2 & F_3 (18 p.)

  f2-3_target Formant measurements taken at relative targets (at the 10% and 90% measurement) intervals using formants F_2 & F_3 (4 p.)

  These tasks were also performed with duration information incorporated as an additional feature.

  f1-2-3_int10_dur Formant measurements taken at 10% intervals using all three formants (28 p.)

  f1-2-3_target_dur Formant measurements taken at relative targets (at the 10% and 90% measurement) using all three formants (7 p.)

  f2-3_int10_dur Formant measurements taken at 10% intervals using formants F_2 & F_3 (19 p.)

  f2-3_target_dur Formant measurements taken at relative targets (at the 10% and 90% measurement) intervals using formants F_2 & F_3 (5 p.)

The methods using parametric representations of formant trajectories as outlined in the previous chapter were split into several categories of tasks along the following dimensions.

First of all, the curves can be fitted to raw formant trajectories (denoted as **raw**) as well as time-equalised (interpolated, denoted **eq**) trajectories. This has been shown to have an impact on the performance (Morrison 2008, 2009b). The functions approximated to raw measurements encode to some extent temporal information in their representation, as the functions are scaled along the abscissa to the length of segment. Hence, considering the additional explicit duration feature, 8 different tasks can be defined per parametric function of each degree, resulting in the following 32 tasks.

- **Polynomial fitting**
  In the following tasks the polynomial fitting method is used on each of the individual segments.

  f1-2-3_poly2_eq Polynomial fitted to time-equalised formant trajectory of all three formants (9 p.)

  f1-2-3_poly2_raw Polynomial fitted to raw formant trajectory of all three formants (9 p.)

  f1-2-3_poly2_dur_eq Polynomial fitted to time-equalised formant trajectory of all three formants, including duration information (10 p.)

  f1-2-3_poly2_dur_raw Polynomial fitted to raw formant trajectory of all three formants, including duration information (10 p.)

  f2-3_poly2_eq Polynomial fitted to time-equalised formant trajectory of the second and third formants (6 p.)

  f2-3_poly2_raw Polynomial fitted to raw formant trajectory of the second and third formants (6 p.)

  f2-3_poly2_dur_eq Polynomial fitted to time-equalised formant trajectory of the second and third formants, including duration information (7 p.)

  f2-3_poly2_dur_raw Polynomial fitted to raw formant trajectory of the second and third formants, including duration information (7 p.)

The tasks using cubic polynomials (**poly3**) and discrete cosine transform curves of second (**dct2**) and third (**dct3**) order are defined in the same manner.

The total number of tasks defined aggregates to 42. As these tasks are performed for the two instances of /aɛ/ in *kreidebleich* both separately and combined, not all task performances can straightforwardly be compared with each other. The relevant comparisons will therefore be picked out and demonstrated, but an exhaustive listing of single-number evaluation results (EER, C_{llr}) for all tasks is provided in tabular form in appendix A.

The tests based on the individual segments are denoted by **aErdB** and **aElCS**, which indicates the diphthong /aɛ/ denoted by *aE*, the immediate context of the segment (r_d in *kreide* and l_ç in *bleich*) and the stress position, with *B* indicating a primary and *S* a secondary stress position. The tests based on all segments combined are simply labelled **aE**.

The results are presented by first looking solely at the discriminatory potential of the methods and parameters chosen and then the calibration properties. The results of applying calibration and fusion techniques are shown at the end of the chapter.

4.3 Results

This section presents the actual results evaluated by the graphical and numerical methods outlined in the previous chapter. The presentation will follow along the following guideline.

- Evaluation and comparison of the performance obtained by different kinds of parametric representations, applied to the individual /aɛ/ segments as well as combined.
- Performance comparison between parametric representations and other methods previously used on diphthongs in forensic speaker recognition

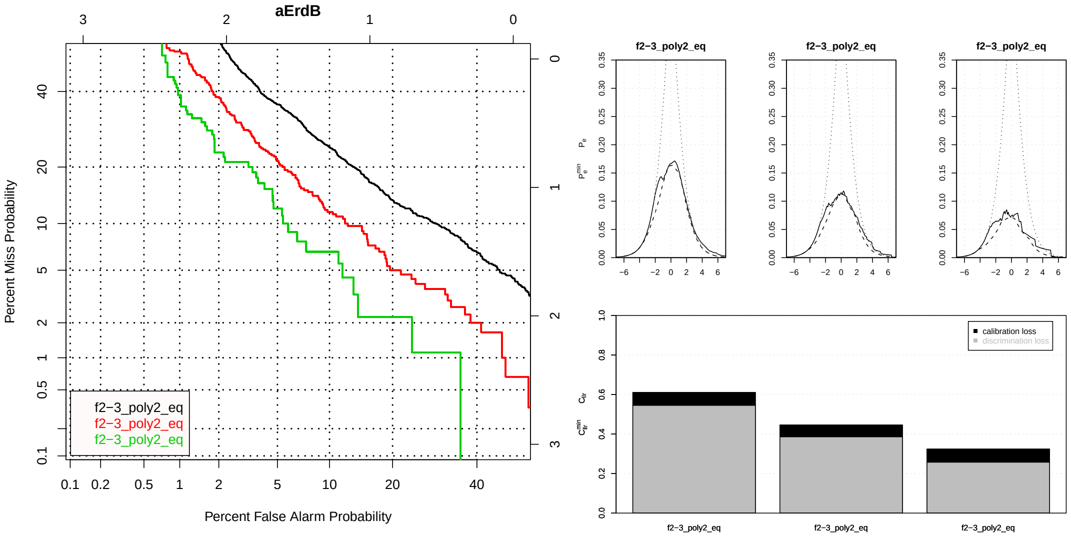
4.3.1 Effect of sample size on evaluation results

Before actual results are shown it is important to note the effects of the sample size. The data available consists of 20 utterances of /aɛ/ of each of the 30 speakers, one half in primary stress position (*kreide*) and the other in secondary stress position (*bleich*), which amounts to 600 tokens in total. The low number of samples per speaker severely limits the options concerning the count of samples used for each trial.

For trials based on segments controlled for same stress position and context there are only ten tokens per speaker. Thus, trials can therefore be constructed using one to ten measurement tokens for modelling the speaker in the likelihood ratio formula. For evaluation purposes, however, target-speaker as well as non-target-speaker trials are needed to obtain meaningful error estimates and confidences.

As the likelihood ratio formula models speakers using Gaussian distributions, a minimum of one measurement is possible, yet not reasonable. Of course, the more measurements are used to represent a speaker the better is the chance of capturing the variability he exhibits in his utterances. The problem, therefore, lies in the question of how to balance the number of measurements and the number of target-trials construed from the data.

Figure 4.3 preliminarily shows the performance of cubic polynomials on /aɛ/ in *kreide* (aErdB) with varying number of measurements and target-trials.



(a) DET plot

(b) APE plot

Figure 4.3: DET and APE plots comparing results with respect to the balance of trials versus measurements

- The black line uses only one measurement for modelling the speaker, thus allows for 45 target trials

- The red line uses two measurements which results in ten target trials
- The green line uses three measurements as well as three target trials

As can be seen there is quite an increase in performance with higher numbers of measurements per trial, yet a sizeable number of target trials is necessary to perform conclusive evaluations of methods. This is also true for a post-hoc calibration stage, where an extensive amount of trials is needed to find good estimates for the parameters used in the calibration procedure.

In the following presentation of results a trade-off has been made to ensure proper evaluation by using two measurements to model a speaker, leading to evaluations based on ten target trials, as well as 145 non-target trials built from the other 29 speakers. This balance is used throughout the present study.

4.3.2 Comparison of parametric representations

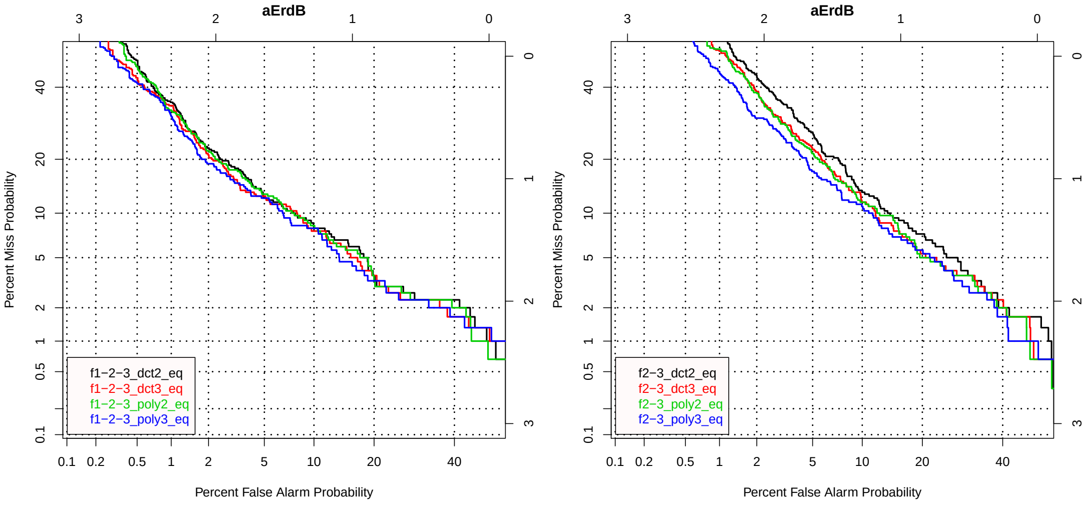
This section presents the results concerning the main topic of the present work, the evaluation of parametric representations. This includes parameters derived from both the polynomial functions and the discrete cosine transform of second and third degree. The performance of these representations was tested under following different conditions.

- Using F_1-F_3 versus using only F_2 & F_3
- Using time-equalised (interpolated) or raw formant trajectories

These conditions were tested on each individual segment (ten utterances per speaker of /aɛ/ in *kreide* and in *bleich*) as well as both pooled together (20 utterances per speaker). The following results deal with the difference in performance in relation to the number of formants used. The two other conditions are investigated in later sections.

Discriminatory potential

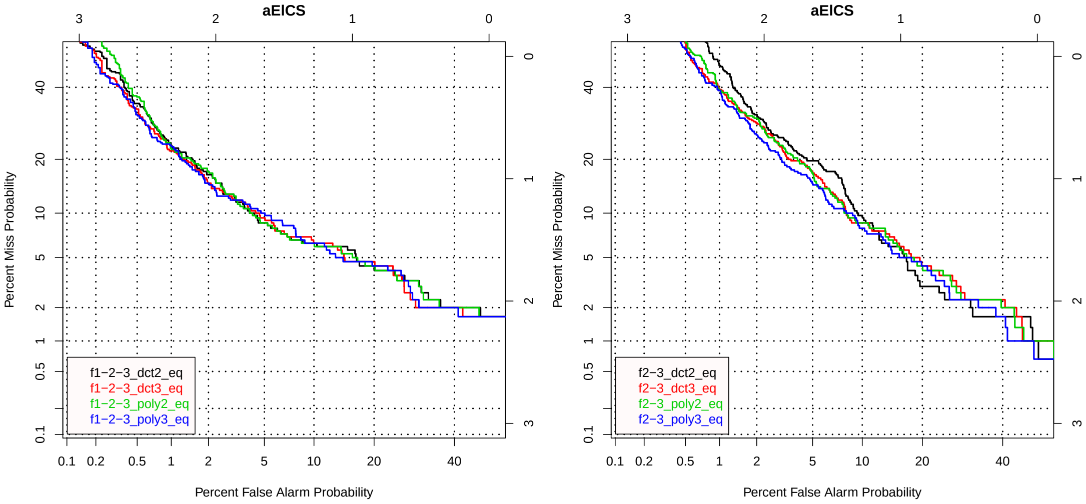
This section presents the comparative results of different parametric representations by means of the DET plot (see section 3.4.4). In figures 4.4 and 4.5 the discriminatory potential of polynomial functions as well as discrete cosine transform curves fitted to time-equalised formant trajectories of the first three formants of both segments are displayed.



(a) F_{1-3}

(b) F_{2-3}

Figure 4.4: DET plot comparing parametric representations of time-equalised formant trajectories in /aɛ/ in *kreide*



(a) F_{1-3}

(b) F_{2-3}

Figure 4.5: DET plot comparing parametric representations of time-equalised formant trajectories in /aɛ/ in *bleich*

As can be seen from the plots useful information can be gained by incorporating the first formant. However, as material obtained from telephone conversations is commonplace in forensic phonetics, the first formant is often compromised by the band-pass filtering and, thus, often cannot be factually used in real casework.

time-equalised			EER		
			aErdB	aElCS	aE
Polynom.	Formants F_1, F_2, F_3	quadratic	0.09	**0.07**	**0.108**
		cubic	**0.083**	0.078	0.11
	Formants F_2, F_3	quadratic	0.11	0.091	0.132
		cubic	0.103	0.093	0.13
DCT	Formants F_1, F_2, F_3	quadratic	0.09	**0.07**	**0.108**
		cubic	0.087	**0.07**	0.113
	Formants F_2, F_3	quadratic	0.121	0.097	0.133
		cubic	0.11	0.087	0.13

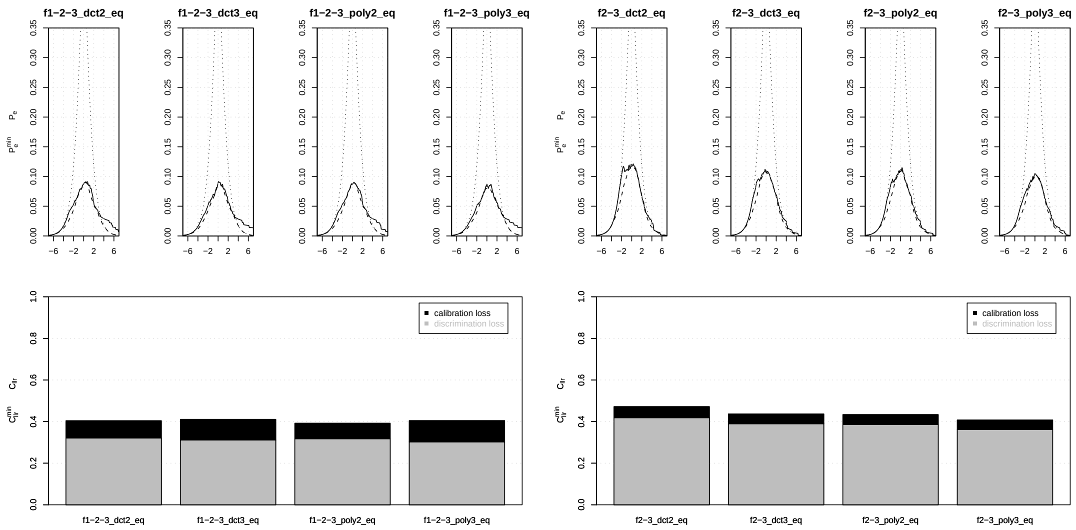
Table 4.1: EER results of polynomial curves fitted to time-equalised formant trajectories

Table 4.1 compares the performance of the curves using EER. Over all segments the difference in performance in terms of absolute EER is only 2% (/aɛ/ in *kreide*) to 2.4% (pooled /aɛ/) in the best-performing tests, i.e. polynomial of second and third degree. Especially noteworthy is the fact that, when comparing the segments against each other, no type of parametric representation stands out as performing best in all cases. The use of the first three formants tested against using only the second and the third delivers systematically better performance, yet no such pattern can be observed with regard to the type of functions. However, there seems to be a slight dominance of polynomial functions, as they generally provide slightly better results.

Calibration performance

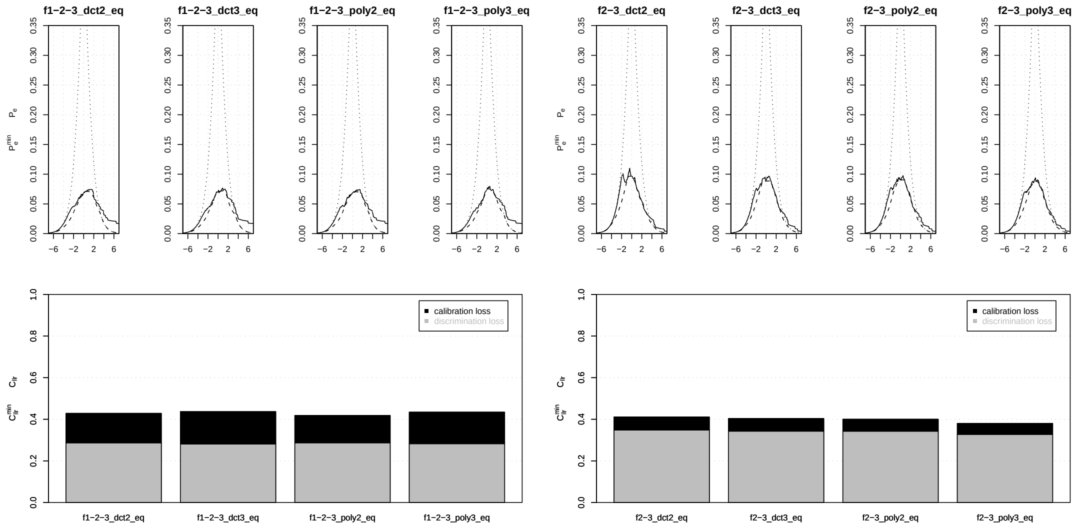
The calibration properties of different parametric representations are compared by means of the APE plot (see section 3.4.4). Figures 4.6 and 4.7 compare the calibration properties of the different parametric representations.

With regards to calibration performance the overall picture is slightly impaired. For /aɛ/ in *kreide* the tests based on using all three formant trajectories exhibit worse calibration properties than those using only F_2 and F_3, however, thanks to their better discriminatory potential their respective C_{llr} values are lower. In terms of this metric, the cubic polynomial representation of the second and third formant achieves as good as the representations derived from all three formants.



(a) F_{1-3} (b) F_{2-3}

Figure 4.6: APE plot comparing parametric representations of time-equalised formant trajectories in /aɛ/ in *kreide*



(a) F_{1-3} (b) F_{2-3}

Figure 4.7: APE plot comparing parametric representations of time-equalised formant trajectories in /aɛ/ in *bleich*

However, for /aɛ/ in *bleich* the C_{llr} values are generally lower in tests using only the second and third formants, although, as indicated by C_{llr}^{min}, their discriminatory potential is indeed higher.

time-equalised			C_{llr}		
			aErdB	aElCS	aE
Polynom.	Formants F_1, F_2, F_3	quadratic	**0.3919**	0.4186	0.4355
		cubic	0.4044	0.4343	**0.4242**
	Formants F_2, F_3	quadratic	0.4334	0.4002	0.5228
		cubic	0.4077	**0.3803**	0.5088
DCT	Formants F_1, F_2, F_3	quadratic	0.4041	0.4286	0.437
		cubic	0.4107	0.4368	0.4332
	Formants F_2, F_3	quadratic	0.4718	0.4112	0.524
		cubic	0.4370	0.4035	0.515

Table 4.2: C_{llr} results of polynomial curves fitted to time-equalised formant trajectories

Table 4.2 compares the performance of the curves using C_{llr}. As with the EER values, no pattern emerges that signalises uniformly better performance in terms of calibration as well as discrimination. However, a slight advantage is exhibited by the polynomial representations over the representations derived from DCT.

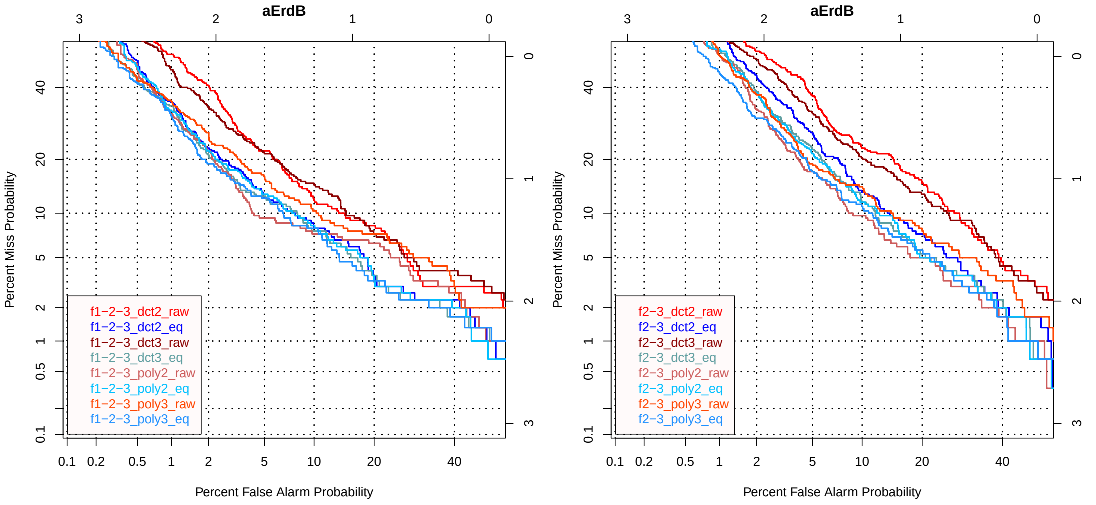
4.3.3 Effect of time-normalisation on performance

One question that arises is the effect of implicit duration modelling which is inherent in approximating parametric functions to formant trajectories as the curves are scaled along the abscissa to the length of the segment when time-normalisation is not applied.

Prior studies (see Morrison (2008, 2009b)) suggest that fitting to time-equalised trajectories shows better performance than when applied to raw formant data. This section investigates this issue and tests if the findings of previous studies hold using Viennese diphthong data.

Figure 4.8 shows comparative DET plots of parametric representations of formant trajectories in /aɛ/ in *kreide*. The left sub-figure shows the performance using data from the first three formants while the right one uses only the second and third formant. The red coloured DET curves represent the tests using raw, non-time-equalised formant trajectories, the blue coloured ones show the performance of parametric representations derived from time-normalised contours.

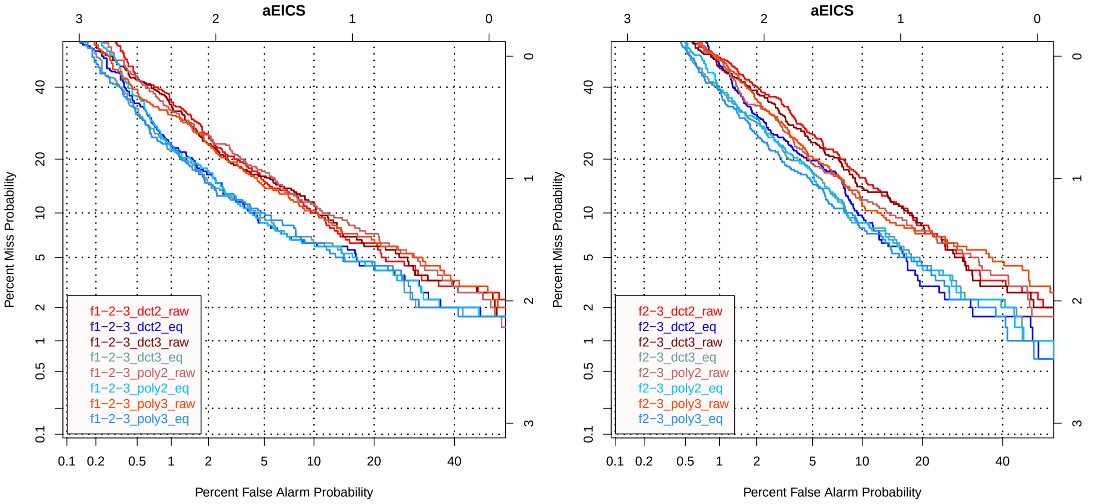
The performance of the methods applied to non-time-equalised data shows extensive spread, suggesting that the parametric functions differ in their ability to generalise over segments of differing length. In this regard the quadratic poly-



(a) F_{1-3} (b) F_{2-3}

Figure 4.8: DET plots comparing parametric representations based on formant trajectories in /aɛ/ in *kreide*

nomial behaves best, surpassing all other parametric functions, and even slightly surpasses the performance exhibited when using time-equalised data.



(a) F_{1-3} (b) F_{2-3}

Figure 4.9: DET plots comparing parametric representations based on formant trajectories in /aɛ/ in *bleich*

Figure 4.9 shows comparative DET plots of parametric representations of formant trajectories in /aɛ/ in *bleich*. As with the previous figure, the methods show better performance when applied to time-normalised data than to raw trajectories. Here, however, the latter tests show less cluttered DET curves, with the curves rather clustered together depending on their respective underlying trajectory data.

4.3.4 Comparison against other approaches

In this section the best-performing parametric representations are compared with other methods applied to the same formant data. As no single parametric function seems to display generally superior behaviour the representations derived from second and third order polynomials fitted to time-equalised trajectories are used to represent the method treated in the present work.

These two are compared to tests based on instantaneous measurements at 10% intervals throughout the formant trajectories (see section 2.2.6, McDougall & Nolan (2007), McDougall (2006)), a dual-target setting emulated by using the 10% as well as the 90% measurement as phonetic targets, and formant means.

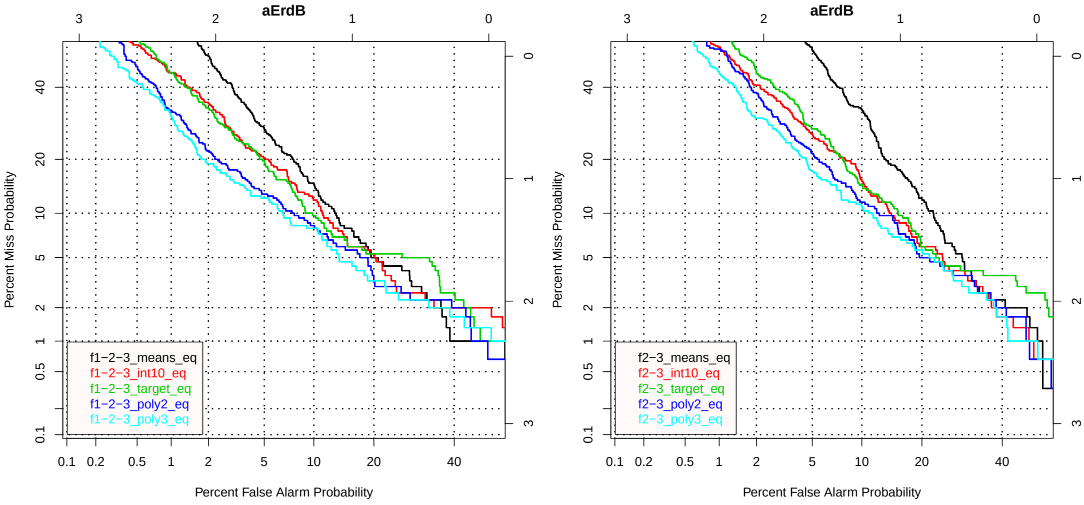
The performance was tested using F_1-F_3 as well as only F_2 & F_3 on each individual segment (10 utterances per speaker of /aɛ/ in *kreide* and in *bleich*) as well as both pooled together (20 utterances per speaker).

Discriminatory potential

As before the discriminatory potential of the methods is first displayed by means of the DET plot. Figures 4.10 and 4.11 compare the five methods previously discussed.

As follows from the DET plots the parametric representations generally perform better than the other methods explored. The measurements at 10% intervals as well as the dual-target tests are the second-best choice, followed by using formant means.

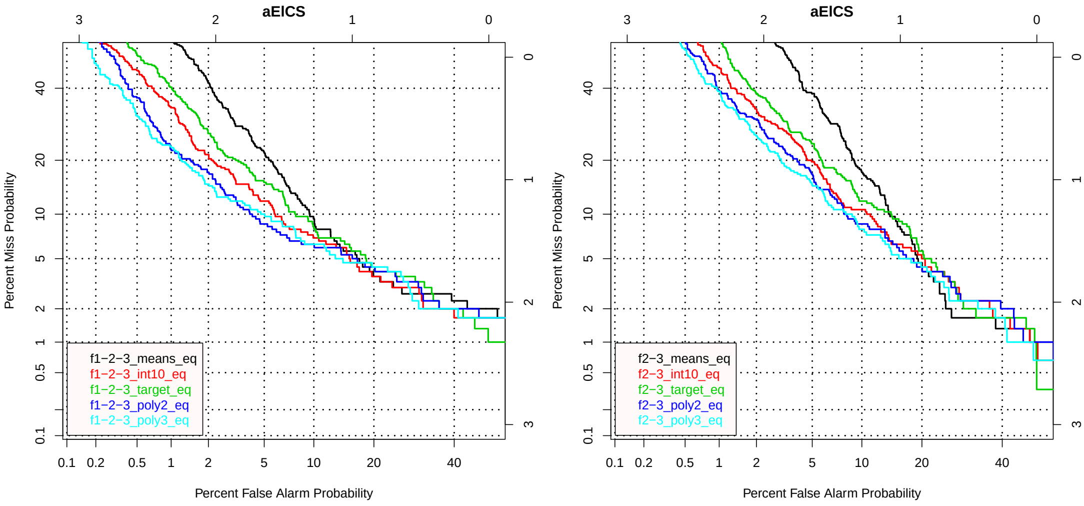
The overall superior performance of the parametric representations holds for both segments and and both the conditions using all three formants or only F_2 & F_3. For a comparison of the individual EER values the same conditions must apply, e.g. the number of formant trajectories used in the tests must be the same.



(a) F_{1-3}

(b) F_{2-3}

Figure 4.10: DET plot comparing parametric representations on time-equalised trajectories with interval, dual-target and means using F_1-F_3 of /aɛ/ in *kreide*



(a) F_{1-3}

(b) F_{2-3}

Figure 4.11: DET plot comparing parametric representations on time-equalised trajectories with interval, dual-target and means using F_1-F_3 of /aɛ/ in *bleich*

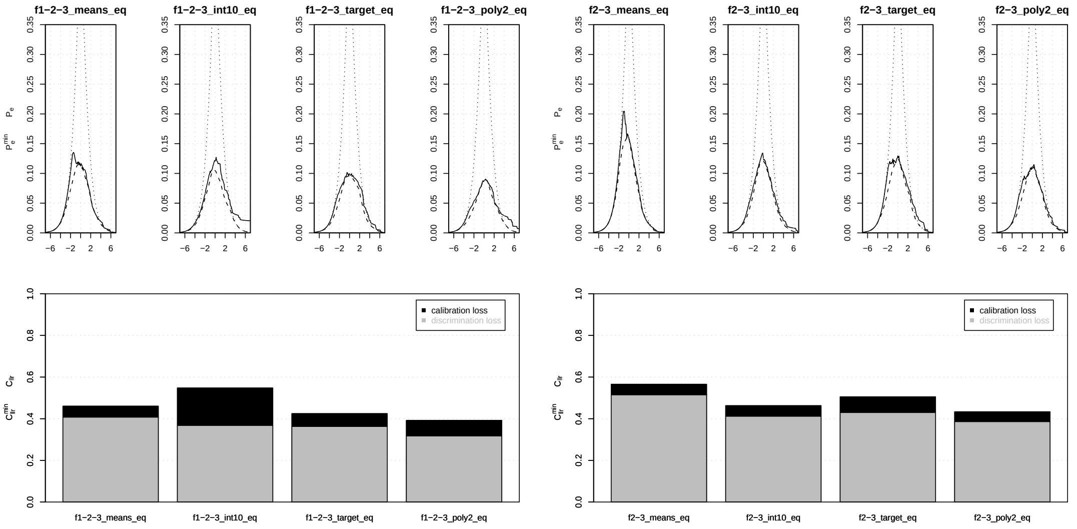
Calibration properties

The calibration properties of the different methods are again compared by means of the APE plot. Figures 4.12 and 4.13 compare the calibration properties of the

			EER		
			aErdB	aElCS	aE
parametric representations	Formants F_1, F_2, F_3	quadratic	0.09	0.07	0.108
		cubic	0.083	0.078	0.11
	Formants F_2, F_3	quadratic	0.11	0.091	0.132
		cubic	0.103	0.093	0.13
instantaneous measurements	Formants F_1, F_2, F_3	10% intervals	0.109	0.08	0.131
		dual-target	0.099	0.093	0.115
	Formants F_2, F_3	10% intervals	0.122	0.105	0.15
		dual-target	0.124	0.114	0.147
formant means	Formants F_1, F_2, F_3		0.115	0.097	0.127
	Formants F_2, F_3		0.165	0.13	0.161

Table 4.3: Comparison of parametric representations on time-equalised trajectories with interval, dual-target and means based on EER values

different parametric representations.

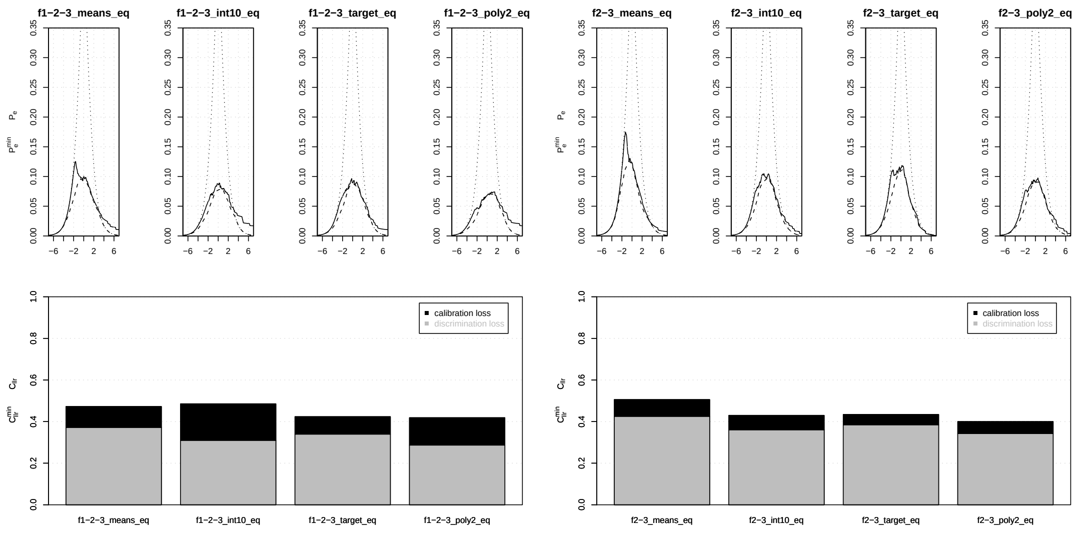


(a) F_{1-3}

(b) F_{2-3}

Figure 4.12: APE plot comparing parametric representations of time-equalised formant trajectories in /aɛ/ in *kreide*

As can be seen the calibration of the parametric representation is good, but this is the case for the other methods too, with the exception of the interval measurements in the condition using all three formants which displays a rather high calibration loss. Another property that becomes visible is that the calibration loss decreases with lower numbers of parameters involved when only using F_2 &



(a) F_{1-3} (b) F_{2-3}

Figure 4.13: APE plot comparing parametric representations of time-equalised formant trajectories in /aɛ/ in *bleich*

			C_{llr}		
			aErdB	aElCS	aE
parametric representations	Formants F_1, F_2, F_3	quadratic	0.3919	0.4186	0.4355
		cubic	0.4044	0.4343	0.4242
	Formants F_2, F_3	quadratic	0.4334	0.4002	0.5228
		cubic	0.4077	0.3803	0.5088
instantaneous measurements	Formants F_1, F_2, F_3	10% intervals	0.5475	0.4854	0.4986
		dual-target	0.4246	0.4237	0.5054
	Formants F_2, F_3	10% intervals	0.4628	0.4299	0.5456
		dual-target	0.5053	0.4343	0.596
formant means	Formants F_1, F_2, F_3		0.4604	0.4725	0.5438
	Formants F_2, F_3		0.5657	0.5058	0.6052

Table 4.4: Comparison of parametric representations on time-equalised trajectories with interval, dual-target and means based on C_{llr} values

F_3, which affects most the methods depending on more features, as with the 10% interval approach which uses 27 parameters for all three formants and 18 for the second and third.

4.3.5 Modelling duration using an explicit parameter

Observing the differences displayed in the time-normalisation condition and given that the results indicate better performance for the parametric representations derived from time-equalised formant trajectories in terms of speaker discrimination than for the ones derived from raw formant contours, the question arises how to incorporate duration information into the speaker model, and how much information could be gained from this procedure.

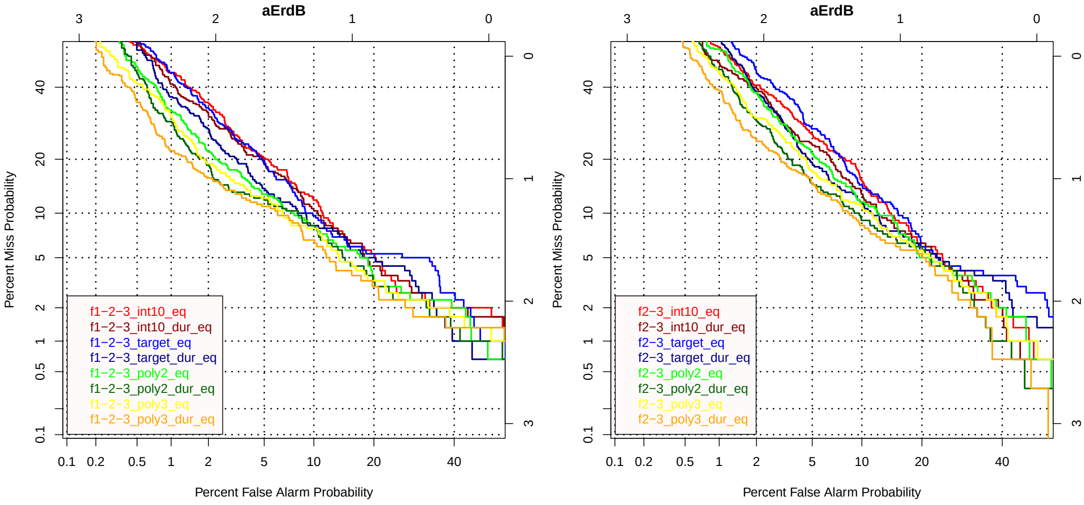
McDougall used the raw duration of the diphthong as another predictor in the discriminant analysis performed in her study, resulting in an improvement of classification rates of 1-5% (McDougall 2005:195). This procedure is applied in the following tests, where the length of the segment is used as another parameter entering the multivariate kernel density formula.

Discrimination performance

Figure 4.14 shows the DET curves for the two best-performing parametric representations, the methods based on instantaneous measurements (10% intervals and dual-target), and the formant means based on the formant trajectories of /aɛ/ in *kreide*. For each method two tests have been made, one including the raw duration in seconds and one in its original form. They are grouped by colors (red, blue, green and yellow/orange).

As can be seen the addition of the duration parameter readily increases the discriminatory potential. The method based on 10% interval measurements shows the lowest increase (EER 0.103 versus 0.109), while the other methods benefited more from the duration information.

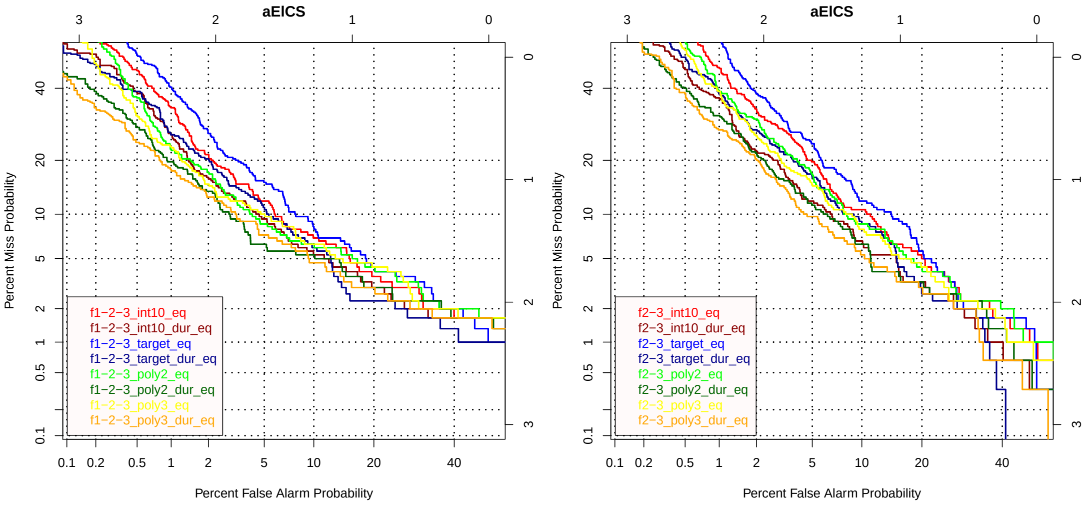
Figure 4.15 shows the performance on data from /aɛ/ in *bleich*. As is apparent from the plot the difference in performance gain by adding the duration parameter is much larger than in the other /aɛ/ segment. The cause of this difference between the two segments predominantly lies in the fact that they are in different stress positions, leading to greater variability in the secondary stressed position in *bleich* than in the diphthong under primary stress in *kreide*. A more thorough account based on Natural Phonology is given in section 5.2.1.



(a) F_{1-3}

(b) F_{2-3}

Figure 4.14: DET plot evaluating the addition of an explicit duration parameter based on formant trajectories of /aɛ/ in *kreide*



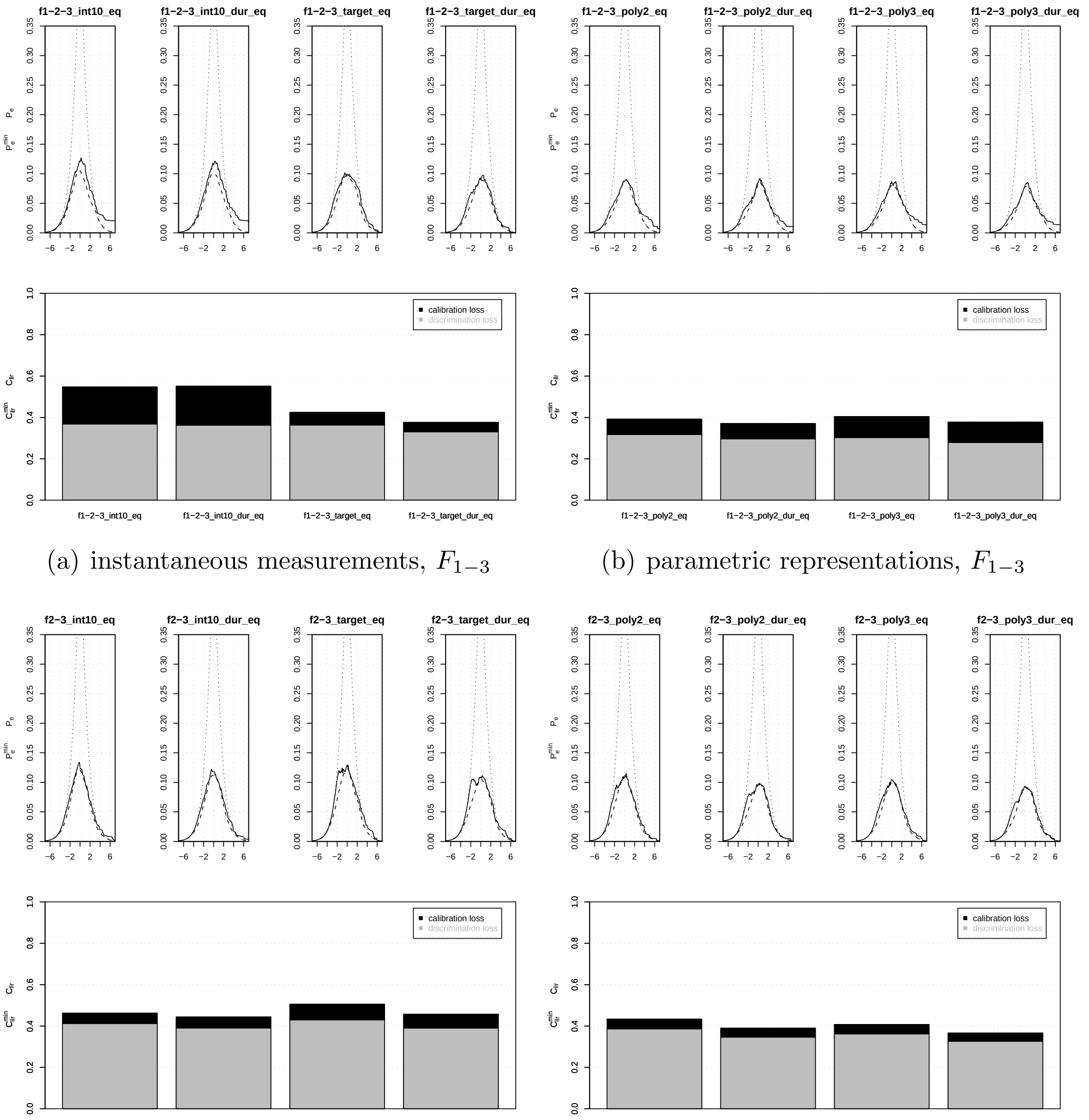
(a) F_{1-3}

(b) F_{2-3}

Figure 4.15: DET plot evaluating the addition of an explicit duration parameter based on formant trajectories of /aɛ/ in *bleich*

Calibration properties

After investigating the potential gain from adding an explicit duration parameter to the model of a speaker the question remains if and how this affects the calibration properties displayed by the different methods. The APE plot as well as C_{llr} is used to quantify the loss due to less-than-optimal LR scale alignment.



(a) instantaneous measurements, F_{1-3}

(b) parametric representations, F_{1-3}

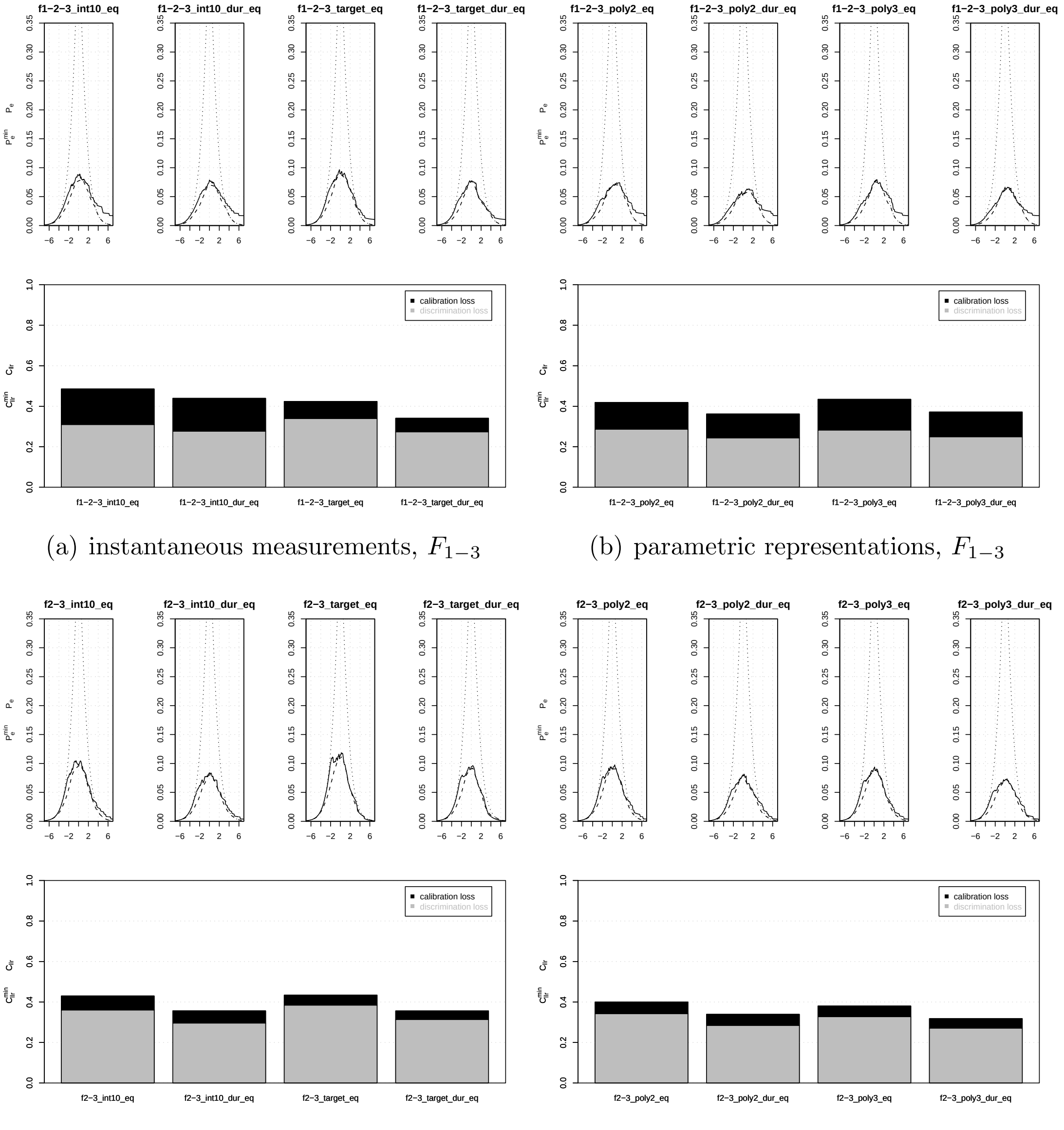
(c) instantaneous measurements, F_{2-3}

(d) parametric representations, F_{2-3}

Figure 4.16: APE plot evaluating the addition of an explicit duration parameter based on formant trajectories of /aɛ/ in *kreide*

Figure 4.16 compares the calibration of the different methods as they were previously used in the other conditions with tests that incorporate the same method as well as the explicit duration feature when applied to /aɛ/ in *kreide*. The pictures to the left display the change in performance of both the instantaneous methods (interval measurements and dual-target) whereas the right side compares the performance of the second and third order polynomials.

As can be seen in the figure the calibration loss indicated by the black section of the bar plot remains more or less constant over all methods and even slightly increases for 10% interval measurements.



(a) instantaneous measurements, F_{1-3}

(b) parametric representations, F_{1-3}

(c) instantaneous measurements, F_{2-3}

(d) parametric representations, F_{2-3}

Figure 4.17: APE plot evaluating the addition of an explicit duration parameter based on formant trajectories of /aɛ/ in *bleich*

Figure 4.17 provides a similar picture in terms of calibration loss, yet clearly displays the gain in discrimination performance achieved when applied to formant data from /aɛ/ in *bleich*. The difference in calibration loss between the conditions using all formants or just the second and third formant are very noticeable.

4.4 Automatic calibration

This section displays the effective gain from applying post-hoc automatic calibration techniques to the likelihood ratios obtained in the tests. The procedure adopted in the present work was laid down in Morrison (2009b), where calibration was performed using cross-validation. This means that for each trial the parameters needed for the calibration stage were trained from the scores of matches and mismatches of all other trials that did not include the speaker (or speakers, in case of non-target trials) involved in the specific trial. This approach was taken to emulate a more realistic picture in that the calibration parameters are estimated from unseen data (see Morrison (2009b:2391)).

In order to show the effect of applying automatic calibration, this section will utilise Tippett plots (see section 3.4.5) as well as the APE plots (section 3.4.7). The presentation is restricted to calibrating the results of polynomials of third degree on time-equalised data based on both individual segments.

Figure 4.18 compares the original performance of cubic polynomials fitted to the trajectories of the first three formants of /aɛ/ in *kreide* with its calibrated counterpart.

As can be seen in the Tippett plot the curve indicating the probability of the defence hypothesis being true saturates much faster than the original curve. The smallest log likelihood ratio value obtained was -87.24 before calibration, which was reduced to -5.05. The extent of reduction results from the parameters specified and obtained by the calibration parameter training stage.

The APE plot shows the reduction in calibration loss, yet indicates a slight loss in discrimination performance. This can also be explained by less-than-optimal calibration parameters, as trials for the training step should be carefully selected for the use in an automatic system to avoid this problem.

Figure 4.19 juxtaposes pre- and post-calibrated likelihood ratios of /aɛ/ in *bleich*. As in the previous plots the Tippett plot shows a very steep curve for the probability of the defence hypothesis being true. Here, the smallest log likelihood

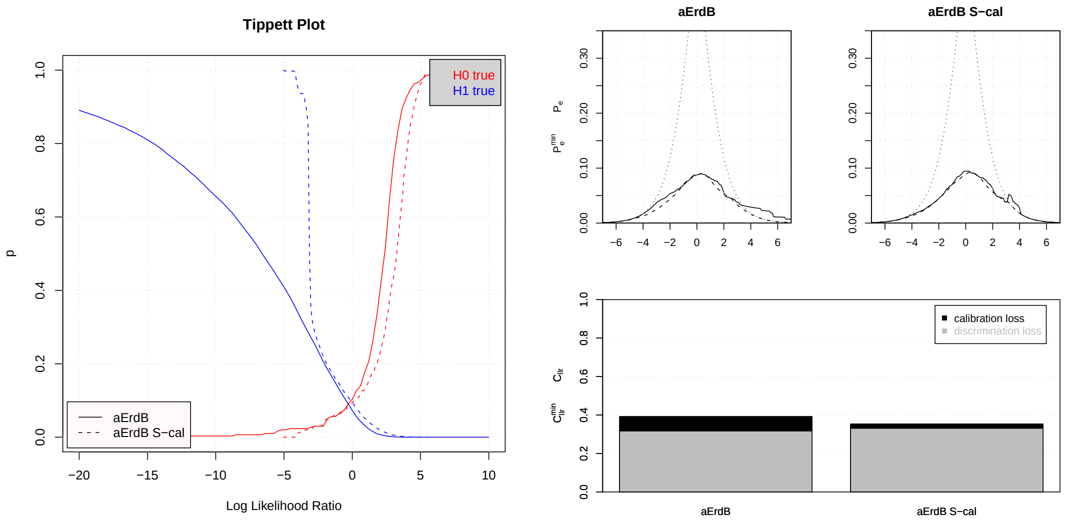


Figure 4.18: Effects of post-hoc calibration on the performance of parametric representation methods based on /aɛ/ in *kreide*

ratio value obtained was -53.8 before calibration, which was reduced to -4.74. Likewise a small loss in discrimination performance can also be observed in the APE plot.

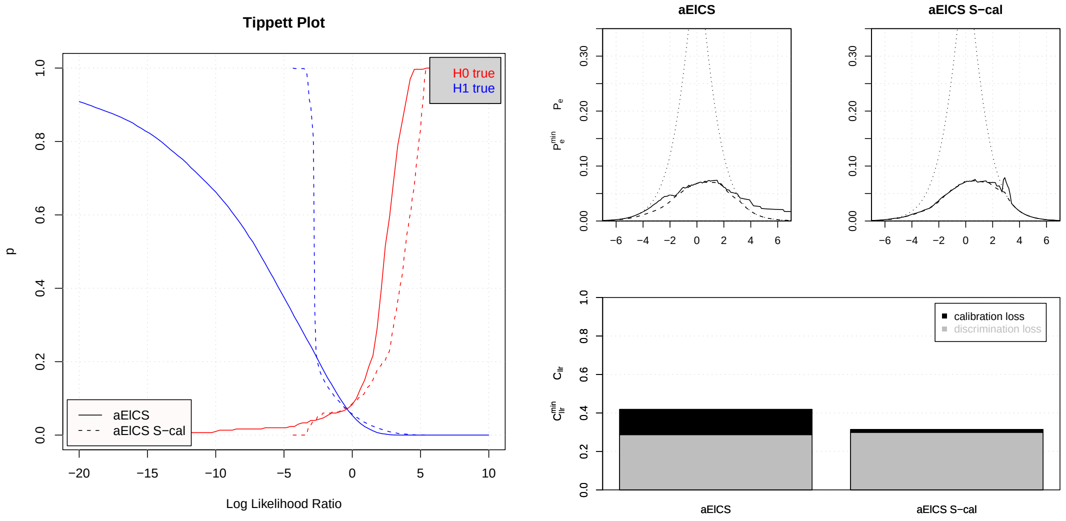


Figure 4.19: Effects of post-hoc calibration on the performance of parametric representation methods based on /aɛ/ in *bleich*

Table 4.5 compares the resulting EER and C_{llr} values before and after the application of the calibration stage for cubic polynomial representations fitted to time-equalised trajectories of the individual segments as well as the pooled segments.

cubic polynomial F_1, F_2, F_3	aErdB		aElCS		aE	
	EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
pre-calibration	0.083	0.4044	0.078	0.4343	0.11	0.4242
post-calibration	0.083	0.3357	0.08	0.3152	0.113	0.4050

Table 4.5: Comparison of EER and C_{llr} before and after calibration

4.5 Automatic fusion of likelihood ratios

This section presents the results obtained by applying the automatic fusion technique based on logistic regression to the discrimination results of the method applied to individual segments. For a detailed description of the procedures involved see section 3.6.

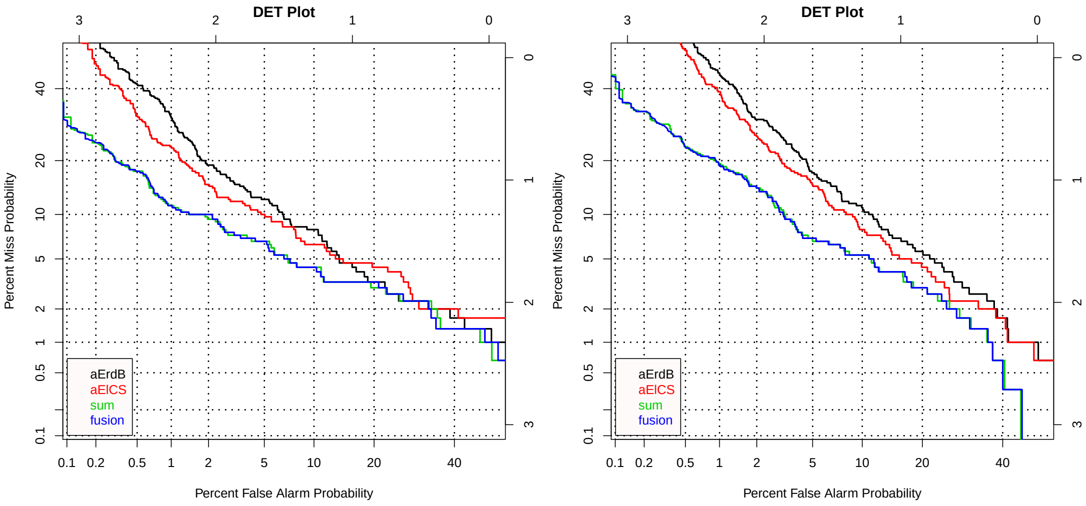
To show the effect of automatic fusion, the likelihood ratio values obtained by this procedure are compared to the sum of the log likelihood ratios of the two segments using the evaluation methods used in the previous sections. The sum of the log likelihood ratios represents the regular method of combining likelihood ratios in the Bayesian approach when independence of the scores of two methods is assumed.

Discriminatory potential

First the added discriminatory potential is considered. Figure 4.20 shows the DET curves produced in the evaluation of performance of the cubic polynomial representation fitted to time-equalised formant trajectories of both individual segments as well as two curves representing the combination of the two segments.

The green line describes the performance of taking the sum of the log likelihood ratios, and the blue line shows the effect of logistic regression fusion.

As can be seen, the two lines representing combination of scores are almost identical. This indicates that the likelihood ratios calculated for the two individual segments tend to agree in their general judgement.



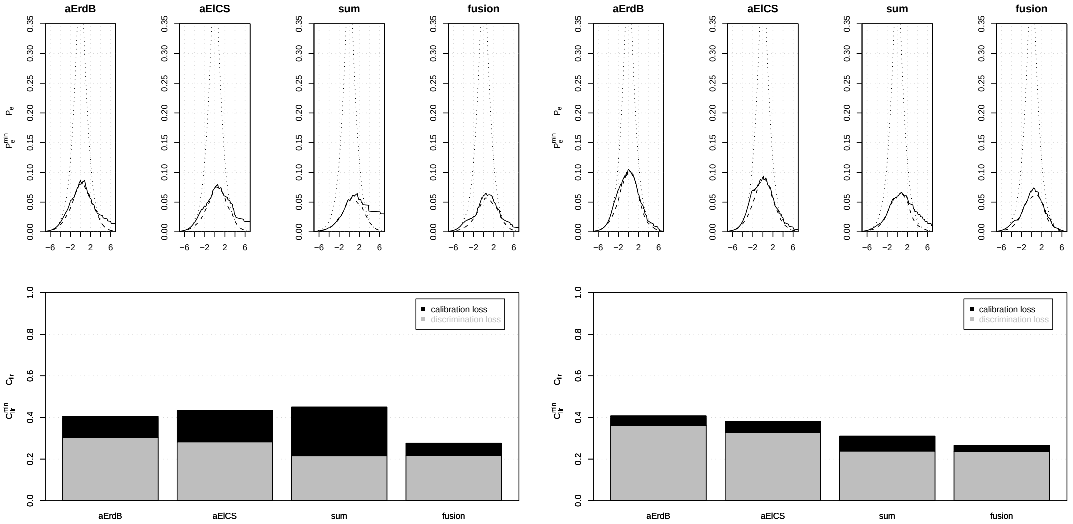
(a) F_{1-3}

(b) F_{2-3}

Figure 4.20: DET plot evaluating the effect of automatic fusion

Calibration properties

As in the previous sections the calibration properties are inspected by means of the APE plot. Figure 4.21 compares the individual segments with their combinations, the log likelihood ratio sum as well as the fused scores.



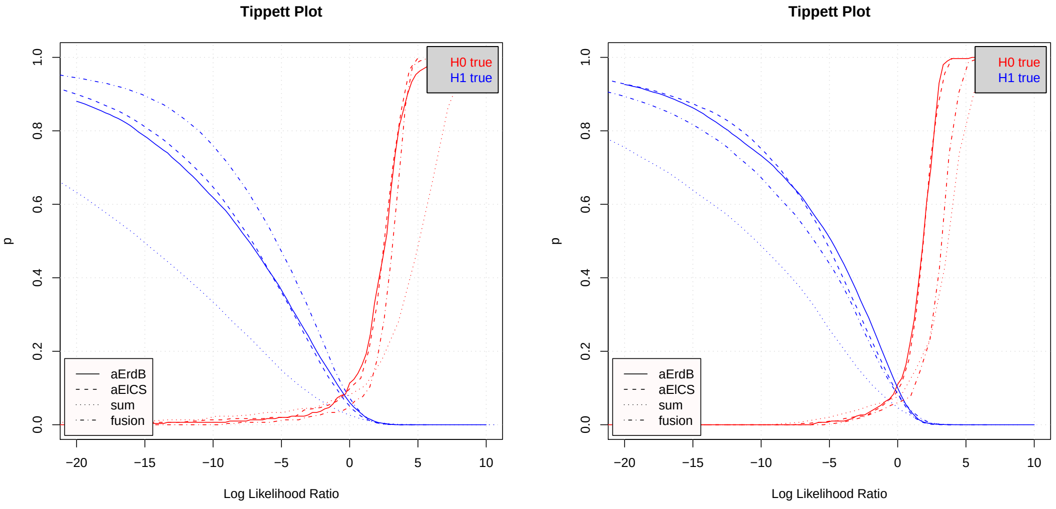
(a) F_{1-3}

(b) F_{2-3}

Figure 4.21: APE plot evaluating the effect of automatic fusion

As can be seen, the automatic calibration stage integrated in the fusion procedure greatly reduces the calibration loss indicated by the black section in the bar plot. The discrimination performance, however, stays the same. The somewhat ample reduction of the calibration loss indicates that the information contributed by the individual segments is quite correlated, yielding astronomically high log likelihood ratios for some trials.

The calibration side-effect of the fusion procedure is also displayed in the Tippett plot in figure 4.22, in which the curve showing the probability of the defence hypothesis being true is much steeper for the fused scores than for the scores combined by addition.



(a) F_{1-3}

(b) F_{2-3}

Figure 4.22: Tippett plot evaluating the effect of automatic fusion

Chapter 5

Discussion

In this chapter the results obtained during the course of the evaluation process are discussed and interpreted. First, a general overview over the performance of the system is given which states general facts concerning the properties of the results. Subsequently, an attempt is made to explain the behaviour of the system depending on characteristics found in the data (section 5.2).

Section 5.3 gives a conclusion of the study presented in this work and finally leads over to section 5.4, which provides for directions for further research.

5.1 General overview

As can be seen from the figures presented in the previous chapter the method described in this study provides consistently better results compared to methods relying on static measurements at particular targets or mean values of the formant values in one trajectory, in terms of its ability to discriminate between speakers as well as of calibration results. This has been expected as previous studies suggested similar improvements in comparison to the other methods, in particular the first study by Morrison (2008) employing polynomial functions as well as a dual-target approach in a likelihood ratio based analysis.

Additionally, the representations derived from polynomial functions generally outperform those from discrete cosine transform (DCT), yet only very slightly. This contrasts to a small extent to the conclusion in Morrison (2009b:2395) which states that '[t]here were trends indicating that the DCTs generally outperformed polynomials [. . .]', although the difference in performance found in the present study is not convincing to signal a contradiction to Morrison's results. Rather, as he himself notes 'the parametric curve with the best performance [...] could only be determined on a case-to-case basis' Morrison (2009b:2395).

The quadratic and cubic polynomials show quite comparable performance, each leading the way in some of the tests. Generally, however, the third order polynomial outperforms the second order polynomial. This can also be observed for the two DCT curves which compares with the results by Morrison (2009b).

The evaluation based on the C_{llr} metric and the Applied Probability of Error (APE) plots show that the likelihood ratio scores delivered by the system are in general well-calibrated in the sense that they are already aligned to the (log) likelihood ratio scale. However, as noted in section 4.4, the values obtained for the non-target-trials reach down to 5.75×10^{-88} before calibration, meaning that they are exorbitantly low, particularly when compared to the numbers calculated for target-speaker trials. A post-hoc calibration stage is therefore advisable to scale the values to a range that is more easily interpretable, as such low and likewise high values can lead to a misleading statement of strength of evidence if they are not handled with care.

Morrison & Kinoshita (2008) note considerable calibration loss in their study of Australian English /oʊ/, with C_{llr} values ranging between 0.6 and more than 1.2. This is in contrast to the results obtained in the present work. Morrison & Kinoshita (2008:1504) state the low number of recordings per speaker as a possible reason. However, the present work is based on only 10 repetitions of *kreidebleich* and therefore uses even less utterances per speaker compared to the study which used 28 recordings of /oʊ/ in different phonemic contexts.

The calibration performed using the sigmoid-based transformation method (*S-cal*, see section 3.5) shows the effect of a post-hoc calibration stage, in which values deviating from the likelihood ratio scale are scaled and shifted, mitigating against possible misinterpretation of the statement given by the score. The method successfully transforms the values down to a range between 8.9×10^{-6} and 3.3×10^{5}.

The conclusions drawn here apply to both sets of tests performed using the first three formant tracks as well as only the second and third formant. Needless to say there is some decline in discrimination ability when using only two formants instead of three, but as outlined in section 4.3.2 the performance loss is quite limited. This is important for the application to recordings taken under less-than-optimal conditions, as it is common in forensic phonetic casework, like for example in telephone conversations and other conditions where the signal is band-pass filtered in any way.

5.2 Interpretation of the results

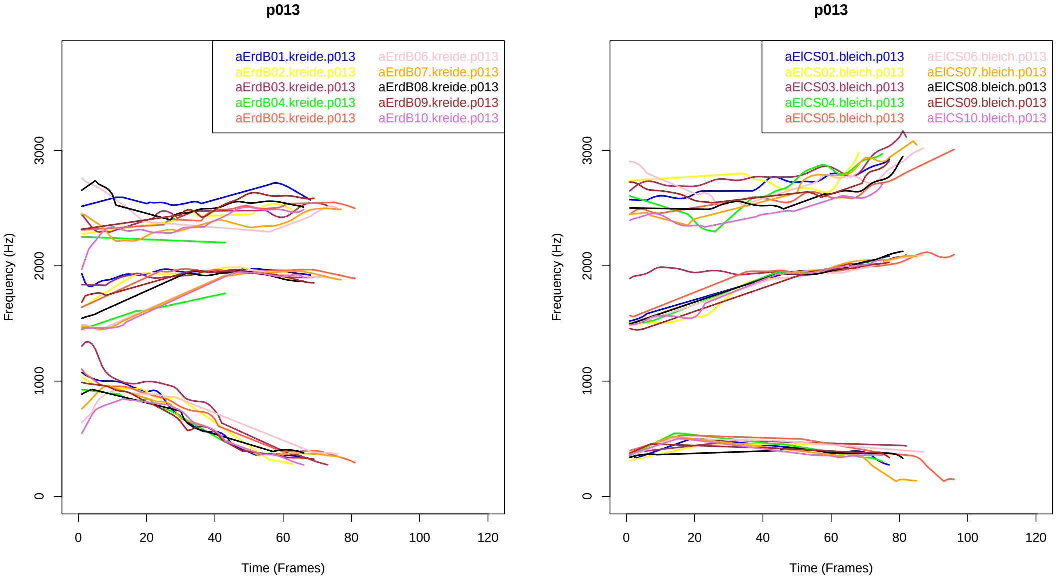
The results show some rather interesting tendencies. First of all, the system performed generally better when applied only to the formant trajectories of the diphthong /aɛ/ in *bleich* instead of in *kreide*. This finding is consistent for each method applied to the data. The actual difference is rather small in terms of error rate, yet, as it is constant over all tests, the cause is likely to be linked with the underlying formant trajectories.

The most apparent differences between the segments are their context and stress positions. The former is especially relevant in this case, as the /r/ phoneme of Viennese German can be realised in several ways. Given the velar plosive context /k/ in *kreide* the most common realisations are an uvular trill [ʀ] or a voiced uvular fricative [ʁ], although alveolar trills [r] are possible as well.

The choice of the particular variant is regarded as dependent on the speaker, the context and the variety or dialect used and therefore is expected to be constant over utterances of one speaker during one recording session. Nevertheless, the /r/ context evidently added variability to the onset of the segment, as explicated in figure 5.1, which shows the raw formant trajectories of both segments for speaker *p013*.

The variation in part (a) of the figure causes a sizeable increase in within-speaker variability quantified by the resulting coefficient values of the fitted curves, while this is not the case for /aɛ/ in *bleich*, shown in part (b). Though there are outliers in the latter too, as well as quite much variability in the third formant, the overall behaviour of the formants is less variable, leading to a safer estimation of models in the speaker discrimination process. This contextual influence could only be reduced by discarding the first part of the trajectory, but the ensuing question about how much should be removed is a delicate one which cannot be answered beforehand.

Returning to stress as the second difference between the two segments, the



(a) /aɛ/ in *kreide*

(b) /aɛ/ in *bleich*

Figure 5.1: Raw formant trajectories of both /aɛ/ segments of speaker p013

monophthongisation process in Viennese German is introduced in the following section, which can be of use to give an account of the increase in performance in terms of error rate when an additional explicit duration parameter is added.

5.2.1 The Viennese monophthongisation process

As has long been noted (Wiesinger 1995:456), there exists a monophthongisation process that is said to have begun around 1900 in Vienna in the speech of the lower social classes which led to a total generalisation. During this process the Standard Austrian German diphthongs /aɛ/ and /ɑɔ/ changed into the monophthongs /æː/ or /ɛː/ and /ɒː/ or /ɔː/ in the Viennese German dialect. Concerning the diphthongs' durational properties, Moosmüller (1997a:787) notes that they 'are said to have been compensated by a lengthening of the resultant monophthongs'.

Acoustic phonetic studies dealing with this process show that 'great variability (from an articulatory point of view) and tolerance (from the point of view of perception) with regard to diphthong articulation can be observed within the Austrian varieties' (Moosmüller 1998:12). In Standard Austrian German only three phonologically relevant diphthongs exist (/aɛ/, /ɔɛ/, and /ɑɔ/). Due to the pervasive nature of the monophthongisation process 'any rising movement in the front or the back vowel space will be interpreted correctly as /aɛ/ or /ɑɔ/ respectively' (Moosmüller 1998:12).

The features resulting from this monophthongisation process are currently spreading to other parts of Austria (Moosmüller 1998:10), which adds to the variability that is to be expected in the general Austrian population.

As Vollmann (1996) points out, a distinction has to be made between the Viennese German dialect and Standard Viennese German which have to be analysed as two independent systems. Speakers of the Viennese Dialect do not produce diphthongs at all due the diachronic development, whereas in Viennese Standard German the monophthongisation seems to be rather gradual (Vollmann 1996). In the dialectal variety all diphthongs historically originating from Middle High German /î/ and /û/ are monophthongised to /ɛː/ and /ɔː/. Those diphthongs tracing to MHG /ei/ and /ou/ turned into /aː/. This distinction, however, does not appear in Standard Viennese German, where they are realised as /aɛ/ and /ɑɔ/ (Moosmüller & Vollmann 2001a:44).

In the framework of Natural Phonology, the relation between dialect and standard can be described by the *two-competence model* (Dressler & Wodak 1982), since all speakers of Austrian German are familiar with both systems and their (socio-)phonological implications (Moosmüller & Vollmann 2001a:43). Interactions between these systems can be accounted for by input-switch rules which refer to opposing forms without being connected by a phonological process. Synchronously there does not exist a relationship between both forms and there are no gradual in-between forms (Moosmüller & Vollmann 2001a:44).

Due to prosodic conditions, monophthongisation can also occur in the Standard Viennese variety. Diphthongs in prosodically weak positions can be produced as monophthongs which, however, are short as compared to dialectal monophthongs in which duration retains it's distinctive role (Moosmüller 1996:1).

The observation that speakers of the Viennese Dialect who are not able to articulate the diphthongs /aɛ/ and /ɑɔ/ use monophthongised forms /ɛː/ and /ɔː/ (Moosmüller & Vollmann 2001a:45) is of particular interest with respect to the results obtained in the tests using time-equalised and raw formant trajectories, as well as in the condition using an additional explicit duration parameter. It was assumed, based on the phonological models set forth in Vollmann (1996) and

Moosmüller & Vollmann (2001b), that the lengthening and thus the duration of the segment was rather invariant for one speaker but highly variable between speakers.

Furthermore, as /aɛ/ in *bleich* is in a prosodically weak position it can be assumed that there are speakers of the Standard variety who produce monophthongised forms that are short in comparison to dialectal realisations. Therefore, the duration can be expected to provide useful additional information to discriminate between users of monophthongs in Viennese German.

5.2.2 Duration differences between segments

As was shown in section 4.3.5, the additional duration parameter resulted in an improvement of performance measured in error rates, which was more substantial for the segment /aɛ/ in *bleich*. To investigate this result the patterns of duration for both segments have to be examined.

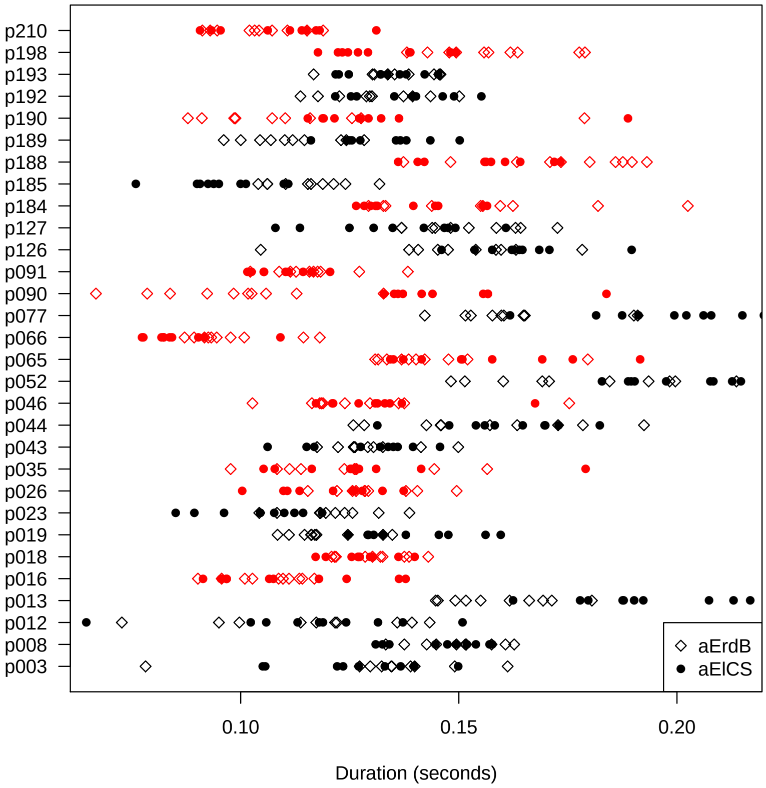
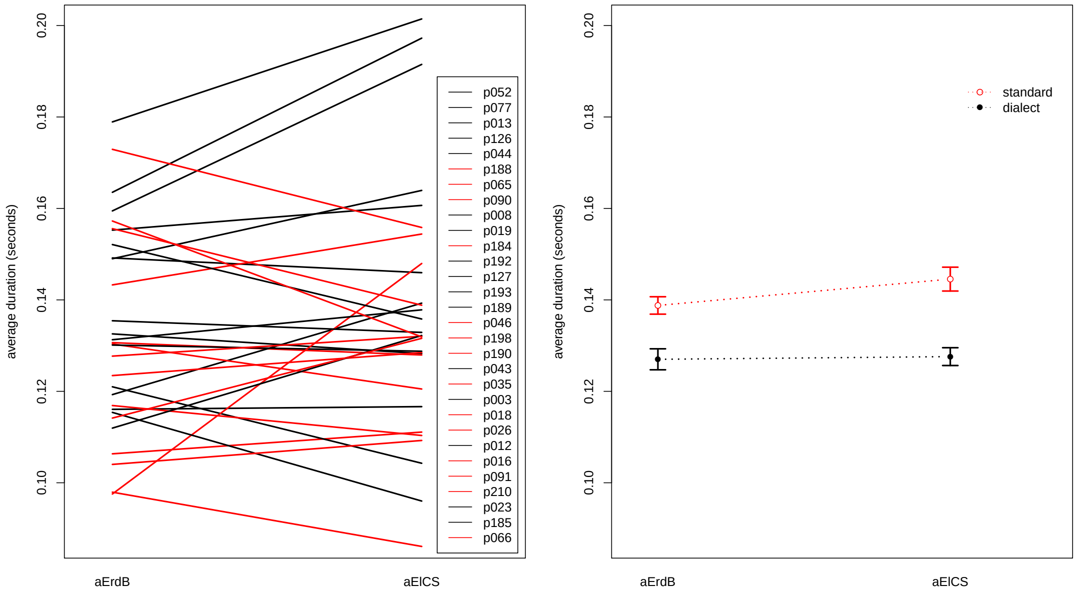


Figure 5.2: Stripchart displaying the durations of both segments for each speaker

Figure 5.2 shows the durations for each speaker in a strip chart, which is basically a one-dimensional scatterplot. Duration measurements of both segments of dialectal and standard speakers are indicated by red and black symbols, respectively. As can be seen there are several speakers who show different duration patterns for each segment.

The question whether there is a difference between duration patterns of dialectal and standard speakers is discussed in the following. Figure 5.3 shows two plots that visualise the interaction between the factor segment and (a) the individual speakers, as well as (b) standard versus dialect. Again, red lines represent the interactions of dialectal speakers and black those of standard speakers. Parallel lines would indicate that there is no variation in average duration between the two segments. However, as can be seen from the plots, there are quite big differences for some speakers, but when comparing averages of dialectal and standard speakers there seems to be only a rather small difference, which suggests that standard speakers produce on average slightly longer /aɛ/ segments in *bleich* than the dialectal speakers, in comparison to the difference in /aɛ/ segments in *kreide*.



(a) individual speakers

(b) standard versus dialect

Figure 5.3: Interaction charts displaying the durations of both segments

To test if there is a statistically significant difference between the durations of the two segments between standard versus dialect speakers, analysis of variance (ANOVA) was performed using a mixed-effects model with standard versus dialect as a between factor and the segment as a within factor, while the speaker was modelled as a factor with random effects. A 5% level of significance was chosen for the test. The F-statistic yielded a value of 5.88 corresponding to a p-value of 0.0156 for the segment as fixed factor, indicating that, at the preassigned level, the null hypothesis of equal average durations between the two segments is rejected. Thus, the difference shown in part (b) of figure 5.3 is significant.

This result suggest that the difference in improvement in terms of error rate when adding an explicit duration parameter can be attributed to the special situation in Viennese German and the monophthongisation process, but further study is needed to check if this outcome also applies to a more general population of Viennese speakers. Concerning the task of speaker discrimination it must be noted, however, that the performance observed in section 4.3.5 must be taken with a grain of salt when the interest is specifically in forensic speaker recognition, as the variability of segment duration within speakers is greatly amplified, for example due to psychological and emotional factors like stress. Thus, it's direct use as a forensic phonetic parameter must be discouraged.

5.3 Conclusion

The method applied in the present work tries to replicate and extend the findings of earlier studies concerning the use of parametric representations of time-dynamic properties of speech as features for the task of forensic phonetic speaker recognition, using data from speakers of Viennese German. As can be seen from the discussion of the evaluation, the results obtained by this method are quite promising, yielding affirming error rates and good calibration properties. In direct comparison with features based on formant means and instantaneous formant measurements at different positions, the two phonetic targets assumed for a diphthong as well as measurements at 10% time intervals throughout the segment, which were chosen as other segmental methods for the evaluation in the present work, it provides a better discriminatory potential and is thus favourable in assessing speaker similarities based on time-dynamic properties within a segment. Various test involving modified feature extraction procedures were performed to assess the influence of difference in trajectory lengths as well as the exclusion of the first formant measurement to emulate conditions like band-pass filtering of the signal.

The expression of the outcome as a likelihood ratio is directed towards the adoption of the Bayesian approach in courts, which is strived for by many practitioners (see Gonzalez-Rodriguez et al. (2007), Rose & Morrison (2009)). It's logical and legal properties have been explained in section 2.3.3. The methods outlined and applied in this work combine traditional forensic phonetic parameters with computerised methods of modelling speaker variability using Aitken & Lucy's likelihood ratio formula. The evaluation and the procedures to calibrate and combine outputs of different systems are incorporated from research on automatic speaker recognition systems in the biometric domain. This emphasises the benefit and need for interdisciplinary research which leads to the utilisation of different types of models and knowledge developed in individual areas in a system which links these techniques to a new application.

The findings presented in this work are in accordance with the results of similar studies by Morrison (2008, 2009b), Morrison & Kinoshita (2008) which applied the method to data from speakers of Australian English, yet are not fully comparable due to the fact that these studies used speech of different recording sessions to account for inter-session variability (see section 2.1.3). However, the true applicability and performance of the method in a forensic phonetic setting can only be assessed when directly applied to real casework data. This study does not satisfy this criterion, as it uses speech recordings that were made under controlled conditions, but it shows the general ability of this approach to discriminate between speakers of Viennese German, given only formant trajectories of diphthongs. The investigation of explicit duration information as an additional parameter must be regarded as a part of this proposition, as it would necessarily provide of less use in realistic forensic conditions.

5.4 Future Research

While the results obtained are indeed promising, extended studies are needed to investigate the performance of the method on other diphthongs as well as monophthongs. Findings like the superiority of the method on data from the secondary stressed diphthong should be taken as an additional incentive to explore other possible aspects in different dialects and sociolects that could affect the performance of these methods. This finding has important implications on its own in that it shows what additional gain in accuracy can be expected of the method given these peculiar circumstances involving specific knowledge of the dialectal situation within the reference population.

To further investigate the applicability of the method and to increase its use in forensic cases to as much material as possible, extended studies will also need to concentrate on the dynamics of other speech sounds than vowels, especially liquids.

As this study is based on studio recordings to test the method's overall applicability, its practical use applied to realistic forensic speech material has yet to be shown. This includes the need for same-speaker recordings taken in different sessions as well as recordings taken under detrimental conditions, like in telephone speech. This last factor is especially noteworthy, as the formant measurements which the method relies on can be severely distorted by band-pass filters, an effect which several studies have shown since (see Künzel, 2001). But, as has been noted in the conclusion, to get insights in the performance of the method for forensic use, the only route is to apply it to real casework data.

As has already been proposed by McDougall (2005:215), future studies should engage in the application of the method to fundamental frequency contours, though less discrimination ability is to be expected due to the very different functions the fundamental frequency fulfils in speech, which, in contrast to formant trajectories, convey non-linguistic information as well.

A last point to mention is the inherent problem of feature correlation not sufficiently treated by this study. The parametric representations applied in the likelihood ratio calculation achieve a great deal of decorrelation which would be present when using several static formant measurements within the trajectory of a diphthong, yet there also exists a correlation between the individual formants incorporated into the analysis. One possible solution to this problem is to adapt the method of likelihood ratio calculation to incorporate prior knowledge about the correlation between these features. This could be achieved by models outlined in Aitken et al. (2006, 2007), where a graphical model estimating the dependency structure is employed to lessen the problem of dimensionality. However, it's applicability to speech data has not yet been shown.

The findings of the present work show that there is much speaker-specific information encoded in the time-dynamic properties of speech segments, which can readily be assumed to be of use when dealing with forensic phonetic casework. This fact is increasingly recognised in recent years and much work has already been devoted to develop methods which exploit this type of information. Yet there is still much potential to be attained by further studies to approach a state of better ability of speaker discrimination.

References

Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Statistics in Practice. John Wiley & Sons.

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

Aitken, C. G. G., Lucy, D., Zadora, G., & Curran, J. (2006). Evaluation of trace evidence for three-level multivariate data with the use of graphical models. *Computational Statististics and Data Analysis*, *50*, 2571–2588.

Aitken, C. G. G., Zadora, G., & Lucy, D. (2007). A two-level model for evidence evaluation. *Journal of Forensic Sciences*, *52(2)*, 412–419.

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

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International*, *146S*, 95–99.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Becker, T. (2008). The influence of intra-speaker variability in automatic speaker verification using f0 features. In *Proceedings of the IAFPA*. Lausanne.

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

Broeders, A. (2001). Forensic speech and audio analysis. forensic linguistics. 1998 to 2001. a review. In *13th INTERPOL Forensic Sciences Symposium*. Lyon, France.

Brümmer, N. (2004). Application-independent evaluation of speaker detection. In *Proceedings of Odyssey-04: The ISCA Speaker and Language Recognition Workshop*. Toledo.

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

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*, 193–203.

Chiba, T., & Kajiyama, M. (1958). *The vowel, its nature and structure*. Phonetic society of Japan.

Donegan, P. J. (2002). Phonological processes and phonetic rules. In K. Dziubalska-Kołaczyk, & J. Weckwerth (Eds.) *Future Challenges for Natural Linguistics*, pp. 57–81. München: Lincom Europa.

Donegan, P. J., & Stampe, D. (1979). The study of natural phonology. In D. A. Dinnsen (Ed.) *Current Approaches to Phonological Theory*. Indiana University Press.

Dressler, W. U. (1984). Explaining natural phonology. In C. J. Ewen, & J. M. Anderson (Eds.) *Phonology Yearbook 1*, vol. 1. Cambridge University Press.

Dressler, W. U., & Wodak, R. (1982). Sociophonological methods in the study of sociolinguistic variation in Viennese German. *Language in Society*, *11*(3), 339–370.

Drygajlo, A. (2007). Forensic Automatic Speaker Recognition. *IEEE Signal Processing Magazine*, *24*(2), 132–135.

Eckert, W. G., & Wright, R. K. (1997). *Introduction to forensic sciences*, chap. Scientific Evidence in Court. CRC Press.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fecher, N. (2008). *Der Einfluss von Telefonsprache auf die Akustik von Vokalen.* Magisterarbeit, Ludwig-Maximilians-Universität, München.

French, J. 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), 137–144.

Gfroerer, S. (2006). *Münchener Anwaltshandbuch Strafverteidigung*, chap. Sprechererkennung und Spurensicherung. G. Widmaier.

Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., & Ortega-Garcia, J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 2104–2115.

Grigoras, C. (2006). Forensic Voice Analysis Based on Long Term Formant Distributions. In *4th European Academy of Forensic Science Conference.*

Grigoras, C., Jessen, M., & Becker, T. (2009). Forensic Speaker Verification Using Long Term Formant Distributions and Likelihood Ratios. In *5th European Academy of Forensic Science Conference.*

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal. *International Journal of Speech, Language and the Law*, *15*(2), 193–218.

Harrington, J. (in press). *The Handbook of Phonetic Sciences*, chap. Acoustic Phonetics. Blackwell.
`http://www.phonetik.uni-muenchen.de/~jmh/research/papers/acoustics2.pdf` (retrieved: 2009-11-02)

Harrington, J., & Cassidy, S. (1999). *Techniques in Speech Acoustics.* Kluwer Academic Publishers.

Holmes, J. N. (2001). *Speech synthesis and recognition.* New York: Taylor & Francis, 2nd ed.

IPA (1999). *Handbook of the International Phonetic Association.* Cambridge, UK: Cambridge University Press.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, *2*, 1–41. Unpublished.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12*(2), 174–213.

Johnson, K. (2003). *Acoustic and Auditory Phonetics*. Cambridge, Massachusetts: Blackwell, 2nd ed.

Joos, M. (1948). Acoustic phonetics. *Language*, *24 (2), Suppl. (= Language Monograph Nr. 23)*.

Keating, P. A. (1990). The window model of coarticulation: articulatory evidence. In J. Kingston, & M. E. Beckman (Eds.) *Papers in laboratory phonology I*, pp. 451–470. Cambridge University Press.

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *International Journal of Speech, Language and the Law*, *12*(2), 235–254.

Kinoshita, Y., & Osanai, T. (2006). Within speaker variation in diphthongal dynamics: what can we compare? In P. Warren, & C. I. Watson (Eds.) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, pp. 112–117. Auckland, New Zealand.

Komsta, L. (2007). *dtt: Discrete Trigonometric Transforms*. R package version 0.1-1.
`http://www.r-project.org,http://www.komsta.net/`

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law*, *8*(1), 80–99.

Ladefoged, P. (2000). *A course in phonetics*. Thomson Wadsworth, 4th edition ed.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35 (11)*, 1773–1781.

Lindh, J. (2004). Preliminary observations on speaker identification in a closed set of disguised voices using ltas (long-time-average-spectrum). In *Proceedings of the IAFPA 2004*. Helsinki.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika*, *64*(2), 207–213.

Markel, J., & Gray, A. (1976). *Linear prediction of speech.* Springer.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, pp. 1895–1898.

McDougall, K. (2005). *The Role of Formant Dynamics in Determining Speaker Identity.* Ph.D. thesis, Department of Linguistics, University of Cambridge.

McDougall, K. (2006). Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, *13.1*, 89–126.

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /uː/ in British English. In J. Trouvain, & W. Barry (Eds.) *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 1825–1828. Saarbrücken.

Moos, A. (2008). *Forensische Sprechererkennung mit der Messmethode LTF (long-term formant distribution).* Magisterarbeit, Universität des Saarlandes.

Moosmüller, S. (1996). The Spread of Viennese Monophthongization: a case of Dialect Levelling?
`http://web.archive.org/web/19971110195930/http://www.esf.org/ftp/pdf/humanities/dialect/s_moosmuller.pdf` (retrieved: 2009-11-03)

Moosmüller, S. (1997a). Diphthongs and the process of monophthongization in Austrian German: A first approach. In *Proc. EUROSPEECH'97*, pp. 787–790. Rhodes, Greece.

Moosmüller, S. (1997b). Phonological variation in speaker identification. *Forensic Linguistics*, *4 (1)*, 29–47.

Moosmüller, S. (1998). The Process of Monophthongization in Austria (Reading Material and Spontaneous Speech). *Papers and Studies in Contrastive Linguistics*, *34*, 9–25.

Moosmüller, S. (2007a). Phonetics needs phonology. In V. A. Vinogradov (Ed.) *Lingvisticeskaja polifonija. Sbornik v cest' jubileja professora R. K. Potapovoj*. Moskau: Jazyki slavjanskich kul'tur.

Moosmüller, S. (2007b). *Vowels in Standard Austrian German. An acoustic-phonetic and phonological analysis*. Habilitationsschrift, Universität Wien.

Moosmüller, S., & Vollmann, R. (2001a). 'Natürliches Driften' im Lautwandel: die Monophthongierung im österreichischen Deutsch. *Zeitschrift für Sprachwissenschaft*, *20/1*, 42–65.

Moosmüller, S., & Vollmann, R. (2001b). The spread of the viennese monophthongization: A socio-phonetic analysis. In *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*. Turin: Rosenberg & Sellier.

Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language and the Law*, *15*(2), 249–266.

Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*, 298–308.

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, *125*(4), 2387–2397.

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 2008 incorporating SST'08*, pp. 1501–1504.

NIST (2008). Speaker Recognition Evaluation Plan.
`http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf` (retrieved: 2009-04-03)

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1997). *The Handbook of Phonetic Sciences*, chap. Speaker Recognition and Forensic Phonetics, pp. 744–767. Blackwell, Oxfored.

Nolan, F. (2001). Speaker identification evidence: its forms, limitations, and roles. In *Proceedings of the conference 'Law and Language: Prospect and Retrospect'*.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, *12*(2), 143–173.

Olsson, J. (2004). *Forensic Linguistics. An introduction to Language, Crime and the Law*. London: continuum.

Przybocki, M. A., Martin, A. F., & Le, A. N. (2007). NIST speaker recognition evaluations utilizing the mixer corpora – 2004, 2005, 2006. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 1951–1959.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
`http://www.R-project.org`

Ramos, D. (2007). *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*. Ph.D. thesis, Universidad Autonoma de Madrid.

Robertson, B., & Vignaux, T. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, chap. Interpreting scientific evidence. Chicester: John Wiley and Sons.
`http://homepages.ecs.vuw.ac.nz/~vignaux/evidence/BookHTML/Ch2Interpretation.htm` (retrieved: 2009-11-03)

Robertson, B., & Vignaux, T. (1997). Bayes' Theorem in the Court of Appeal. *The Criminal Lawyer*, *January*, 4–5.
`http://homepages.mcs.vuw.ac.nz/~vignaux/docs/logicalJuries.html` (retrieved: 2009-11-12)

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

Rose, P. (2005). Forensic speaker recognition at the beginning of the twenty-first century - an overview and a demonstration. *Australian Journal of Forensic Sciences*, *37*, 49–72.

Rose, P. (2006a). The intrinsic forensic discriminatory power of diphthongs. In *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*, pp. 64–69.

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

Rose, P., Kinoshita, Y., & Alderman, T. (2006). Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. In *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*, pp. 329–334.

Rose, P., & Morrison, G. S. (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech, Language and the Law*, *16*(1), 139–163.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3–45.

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

van Leeuwen, D. A., & Brümmer, N. (2007). *Speaker Classification I. Fundamentals, Features, and Methods*, chap. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, pp. 330–353. Springer.

Vollmann, R. (1996). Phonetics of informal speech: The Viennese Monophthongization. *Studia Phonetica Posnaniensia*, *5*, 87–100. `http://www.kfunigraz.ac.at/~vollmanr/pubs/VR1996C.html` (retrieved: 2009-04-15)

Watson, C., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, *106*, 458–468.

Whalen, D. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*, 3–35.

Wiesinger, P. (1995). Varietäten der gegenwärtigen Wiener Stadtsprache. Gebrauch–Einschätzung–Wandel. In G. Lerchner, M. Schröder, & U. Fix (Eds.) *Chronologische, areale und situative Varietäten des Deutschen in der Sprachhistoriographie. Festschrift für Rudolf Große*, pp. 447–460. Frankfurt am Main: Peter Lang.

Xue, S. A., & Hao, J. G. (2006). Normative Standards for Vocal Tract Dimensions by Race as Measured by Acoustic Pharyngometry. *Journal of Voice*, *20*(3), 391–400.

Appendix A

Task performance results

A.1 Formant means

	aErdB		aElCS		aE	
	EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Formants F_1, F_2, F_3	0.115	0.4604	0.097	0.4725	0.127	0.5438
Formants F_2, F_3	0.165	0.5657	0.13	0.5058	0.161	0.6052

Table A.1: Results based on formant mean values

A.2 Interval and target measurements

			aErdB		aElCS		aE	
			EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Formants F_1, F_2, F_3	dur	10% intervals	0.103	0.5516	0.07	0.4392	0.12	0.4616
		Dual-Target	0.091	0.3763	0.076	0.3408	0.109	0.4551
		10% intervals	0.109	0.5475	0.08	0.4854	0.131	0.4986
		Dual-Target	0.099	0.4246	0.093	0.4237	0.115	0.5054
Formants F_2, F_3	dur	10% intervals	0.113	0.4441	0.083	0.3564	0.127	0.4904
		Dual-Target	0.107	0.4569	0.096	0.3561	0.129	0.529
		10% intervals	0.122	0.4628	0.105	0.4299	0.15	0.5456
		Dual-Target	0.124	0.5053	0.114	0.4343	0.147	0.596

Table A.2: Results based on instantaneous measurements (10% intervals, dual-target)

A.3 Parametric representations

A.3.1 Polynomial functions

time-equalised			aErdB		aElCS		aE	
			EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Formants F_1, F_2, F_3	dur	quadratic	0.087	0.3708	0.057	0.362	0.103	0.4015
		cubic	0.083	0.3775	0.065	0.3717	0.099	0.3877
		quadratic	0.09	0.3919	0.07	0.4186	0.108	0.4355
		cubic	0.083	0.4044	0.078	0.4343	0.11	0.4242
Formants F_2, F_3	dur	quadratic	0.097	0.3896	0.08	0.3392	0.119	0.4689
		cubic	0.093	0.3662	0.07	0.318	0.115	0.4482
		quadratic	0.11	0.4334	0.091	0.4002	0.132	0.5228
		cubic	0.103	0.4077	0.093	0.3803	0.13	0.5088

Table A.3: Results of polynomial curves fitted to time-equalised formant trajectories

raw trajectories			aErdB		aElCS		aE	
			EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Formants F_1, F_2, F_3	dur	quadratic	0.087	0.4153	0.09	0.4242	0.121	0.4825
		cubic	0.097	0.4833	0.093	0.4653	0.123	0.5119
		quadratic	0.083	0.4132	0.107	0.4932	0.13	0.5303
		cubic	0.103	0.4961	0.103	0.5437	0.14	0.5617
Formants F_2, F_3	dur	quadratic	0.087	0.3699	0.102	0.3932	0.128	0.5028
		cubic	0.093	0.405	0.097	0.4859	0.125	0.5095
		quadratic	0.097	0.4023	0.114	0.48	0.147	0.5993
		cubic	0.118	0.4627	0.104	0.5849	0.144	0.6057

Table A.4: Results of polynomial curves fitted to raw, non-time-equalised formant trajectories

A.3.2 Discrete cosine transform (DCT)

time-equalised			aErdB		aElCS		aE	
			EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Formants F_1, F_2, F_3	dur	2nd order	0.089	0.3828	0.057	0.3706	0.101	0.4018
		3rd order	0.086	0.3803	0.063	0.3714	0.1	0.3953
		2nd order	0.09	0.4041	0.07	0.4286	0.108	0.437
		3rd order	0.087	0.4107	0.07	0.4368	0.113	0.4332
Formants F_2, F_3	dur	2nd order	0.093	0.3944	0.083	0.3431	0.121	0.4691
		3rd order	0.093	0.3744	0.07	0.3221	0.117	0.4556
		2nd order	0.121	0.4718	0.097	0.4112	0.133	0.524
		3rd order	0.11	0.4370	0.087	0.4035	0.13	0.515

Table A.5: Results of discrete cosine transform (DCT) representations derived from time-equalised formant trajectories

raw trajectories			aErdB		aElCS		aE	
			EER	C_{llr}	EER	C_{llr}	EER	C_{llr}
Formants F_1, F_2, F_3	dur	2nd order	0.118	0.5083	0.097	0.4262	0.147	0.5484
		3rd order	0.13	0.5411	0.102	0.4321	0.145	0.5386
		2nd order	0.113	0.4942	0.1	0.4414	0.143	0.5411
		3rd order	0.127	0.5123	0.102	0.4477	0.143	0.5348
Formants F_2, F_3	dur	2nd order	0.17	0.5712	0.133	0.496	0.167	0.6278
		3rd order	0.15	0.5635	0.131	0.484	0.216	0.6851
		2nd order	0.17	0.5766	0.128	0.5148	0.171	0.6405
		3rd order	0.157	0.5738	0.133	0.5006	0.17	0.6362

Table A.6: Results of discrete cosine transform (DCT) representations derived from raw, non-time-equalised formant trajectories

Abstract

English

The present work investigates the performance of an approach for forensic speaker recognition that is based on parametric representations of formant trajectories. Quadratic and cubic polynomial functions are fitted to formant contours of diphthongs. The resulting coefficients as well as the first three to four components derived from discrete cosine transform (DCT) are used in order to capture the dynamic properties of the underlying speech acoustics, and thus of the speaker characteristics. This results in a representation based on only a small number of decorrelated parameters that are in turn used for forensic speaker recognition. The evaluation conducted in the study incorporates the calculation of likelihood ratios for use in the Bayesian approach of evidence evaluation. The advantages of this framework and its current limitations are discussed.

For the calculation of the likelihood ratios a multivariate kernel density formula developed by Aitken & Lucy (2004) is used which takes both between-speaker and within-speaker variability into account. Automatic calibration and fusion techniques as they are used in automatic speaker identification systems are applied to the resulting scores. To further investigate the importance of duration aspects of the diphthongs for speaker recognition an experiment is undertaken that evaluates the effect of time-normalisation as well as modelling segment durations using an explicit parameter. The performance of the parametric representation approach compared with other methods as well as the effects of calibration and fusion are evaluated using standard evaluation tools like the detection error trade-off (DET) plots, the applied probability of error (APE) plot, the Tippett plot as well as numerical indices like the EER and the C_{llr} metric.

Deutsch

Die vorliegende Arbeit untersucht das Leistungsverhalten eines Ansatzes der forensischen Sprechererkennung, der auf parametrischen Repräsentationen von Formantverläufen basiert. Quadratische und kubische Polynomfunktionen werden dabei an Formantverläufe von Diphthongen angenähert. Die resultierenden Koeffizienten sowie die ersten drei bzw. vier Komponenten der Diskreten Kosinustransformation (DCT) werden in Folge verwendet, um die dynamischen Eigenschaften der zugrundeliegenden akustischen Merkmale der Sprache und damit der Sprechercharakteristika zu erfassen. Am Ende steht eine Repräsentation bestehend aus wenigen dekorrelierten Parametern, die für die forensische Sprechererkennung verwendet werden. Die in der Untersuchung durchgeführte Evaluierung beinhaltet die Berechnung von Likelihood-Ratio-Werten für die Anwendung im Bayesschen Ansatz für die Bewertung von forensischen Beweisstücken. Die Vorteile dieses Systems und die derzeitigen Beschränkungen werden behandelt.

Für die Berechnung der Likelihood-Ratio-Werte wird eine von Aitken & Lucy (2004) entwickelte multivariate Kernel-Density-Formel verwendet, die sowohl Zwischen-Sprecher- als auch Inner-Sprecher-Variabilität berücksichtigt. Automatische Kalibrierungs- und Fusionstechniken, wie sie in Systemen zur automatischen Sprecheridentifikation verwendet werden, werden auf die Ergebniswerte angewendet.

Um die Bedeutung von Längenaspekten von Diphthongen für die forensische Sprechererkennung näher zu untersuchen wird ein Experiment durchgeführt, in dem der Effekt von Zeitnormalisierung sowie die Modellierung der Dauer durch einen expliziten Parameter evaluiert werden.

Die Leistungsfähigkeit der parametrischen Repräsentationen verglichen mit anderen Methoden sowie die Effekte der Kalibrierung und Fusion werden unter Verwendung üblicher Bewertungswerkzeuge wie des Erkennungsfehlerabwägungs-(DET)-Diagramms, des Tippett-Diagramms und des angewandten Fehlerwahrscheinlichkeits-(APE)-Diagramms, sowie numerischer Kennziffern wie der Gleichfehlerrate (EER) und der C_{llr}-Metrik evaluiert.

Curriculum vitae

Persönliche Daten

Name	Ewald Enzinger
Geburtsdatum	27. 09. 1984
Geburtsort	St. Pölten
Staatsangehörigkeit	Österreich

Bisherige Ausbildung

1991 – 1995	Volksschule Neulengbach
1995 – 1999	Bundesgymnasium St. Pölten
1999 – 2004	Höhere Technische Bundeslehr- und Versuchsanstalt, Abteilung für EDV und Organisation, St. Pölten
2007 – 2008	ERASMUS Auslandssemester, Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart
2005 – 2009	Universität Wien, Diplomstudium Allgemeine und Angewandte Sprachwissenschaft, Schwerpunkt Computerlinguistik, Wien

Studienbezogene Tätigkeiten

08/08 – 10/08	Institut für Schallforschung, Österreichische Akademie der Wissenschaften, *Praktikum*
11/08 – 04/09	Institut für Schallforschung, Österreichische Akademie der Wissenschaften, *Wissenschaftlicher Mitarbeiter, Abteilung Software*
05/09	Institut für Schallforschung, Österreichische Akademie der Wissenschaften, *Wissenschaftlicher Mitarbeiter, Phonetik/Digitale Signalverarbeitung*